Department Copy

Generalized Additive Models

Trevor Hastie

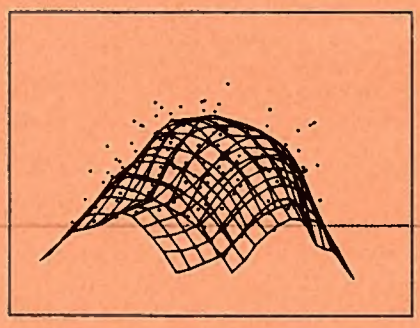
and

Robert Tibshirani

Technical Report No. 2

September 1984

Laboratory for Computational Statistics



Department of Statistics
Stanford University

Generalized Additive Models

Trevor Hastie,

and

Robert Tibshirani

both of

Department of Statistics

Stanford University

and

Computation Group

Stanford Linear Accelerator Center

Abstract

Likelihood based regresssion models, such as the normal linear regression model and the linear logistic model, assume a linear (or some other parametric) form for the covariate effects. We introduce the *Local Scoring* procedure which replaces the linear form $\sum X_j\beta_j$ by a sum of smooth functions $\sum s_j(X_j)$. The $s_j(\cdot)$'s are unspecified functions that are estimated using scatterplot smoothers. The technique is applicable to any likelihood-based regression model: the class of *Generalized Linear Models* contains many of these. In this class the *Local Scoring* procedure replaces the linear predictor $\eta = \sum X_j\beta_j$ by the additive predictor $\sum s_j(X_j)$; hence the name *Generalized Additive Models*. Local Scoring can also be applied to non-standard models like Cox's proportional hazards model for survival data.

In a number of real data examples, the Local Scoring procedure proves to be useful in uncovering non-linear covariate effects. It has the advantage of being completely automatic, i.e no "detective work" is needed on the part of the statistician.

In a further generalization, the technique is modified to estimate the form of the *link function* for generalized linear models.

The Local Scoring procedure is shown to be asymptotically equivalent to *Local Likelihood* estimation, another technique for estimating smooth covariate functions. They are seen to produce very similar results with real data, with Local Scoring being considerably faster.

As a theoretical underpinning, we view Local Scoring and Local Likelihood as empirical maximizers of the *expected log-likelihood*, and this makes clear their connection to standard maximum likelihood estimation.

A method for estimating the "degrees of freedom" of the procedures is also given.

---

* This work was supported by the Department of Energy under contracts DE-AC03-76SF and DE-AT03-81-ER10843, and by the Office of Naval Research under contract ONR N00014-81-K-0340, and by the U.S. Army Research Office under contract DAAG29-82-K-0056.

1. Introduction.

Likelihood-based regression models are important tools in data analysis. A typical scenario goes as follows. A likelihood is assumed for a response variable Y, and the mean or some other parameter is modeled as a linear function of a set of covariates $X_1, X_2, \ldots X_p$. The parameters of the linear function are then estimated by maximum likelihood. Examples of this are the normal linear regression model, the logistic regression model for binary data, and Cox's proportional hazards model for survival data. These models all assume a linear (or some parametric) form for the covariates effects.

In the linear regression problem, there has been a trend in the past few years to move away from linear functions and model the dependence of Y on $X_1, X_2, \ldots X_p$ in a more non-parametric fashion. For a single covariate, such a model would be $Y = s(X) + error$ where $s(X)$ is a unspecified smooth function. This function can be estimated by any so-called *scatterplot smoother*, for example a running mean, running median, running least squares line, kernel estimate, or a spline. The resulting smooth estimate is useful both as a descriptive tool and as a predictive model. For the p covariates $X_1, X_2 \ldots X_p$, one can use a k-dimensional scatterplot smoother to estimate $s(\boldsymbol{X})$, or else assume a less general model such as $s(\boldsymbol{X}) = \sum_1^p s_j(X_j)$ and estimate it in a forward stepwise manner.

In this paper, we propose a technique for estimating smooth covariate functions in any likelihood-based regression model. We call it the *Local Scoring* algorithm. This procedure uses scatterplot smoothers to generalize the usual Fisher Scoring procedure for computing maximum likelihood estimates. For example, the linear logistic model for binary data specifies $\log p/(1-p) = \beta_0 + \beta_1 X_1 + \cdots \beta X_p$. This can be generalized to $\log p/(1-p) = \sum_1^p s_j(X_j)$, and the *Local Scoring* procedure provides non-parametric, smooth estimates of the $s_j(\cdot)$'s.

The Gaussian and logistic models are members of the class of *generalized linear models* (GLM's) (Nelder and Wedderburn, 1972). This comprehensive class restricts Y to be in the exponential family; the statistical package GLIM (Generalized Linear Interactive Modeling) performs estimation and diagnostic checking of these models. The local scoring procedure generalizes GLM's by replacing the linear predictor $\eta = \sum X_j \beta_j$ by an additive predictor of the form $\eta = \sum s_j(X_j)$. The third example mentioned earlier, the proportional hazards model, is not in the exponential family, and the likelihood it uses is not in fact a true likelihood at all. Nevertheless, we still think of it as a "likelihood-based" regression model, and the local scoring procedure can be applied. The usual form for the relative risk, $\exp(\sum X_j \beta_j)$ is replaced by the more general form $\exp(\sum s_j(X_j))$.

The local scoring procedure is asymptotically equivalent to another method for estimating smooth covariate functions, *Local Likelihood* estimation (Tibshirani, 1982, Hastie, 1983a, and Tibshirani, 1984). In this paper we compare the two techniques in some examples and find that the estimated functions are very similar. The advantage of the local scoring method is that it is considerably faster.

As a further generalization, the local scoring procedure can be extended to provide non-parametric estimation of the link function. This facilitates, for example, estimation of the model $f(p) = \sum s_j(X_j)$ for binary data. It also provides a check of two of the assumptions inherent in linear logistic modeling: the linear form for the covariates and the logit link. The complete algorithm, with covariate and link function estimation, can be viewed as a generalization of Friedman and Owen's PACE model to non-Gaussian data (Friedman and Owen, 1984).

This paper is non-technical for the most part, with an emphasis on the techniques and their illustration through examples. In Section 2, we review the linear regression model and its generalization (additive models built up from covariate smooths). Section 3 reviews generalized linear models. In Section 4, we link smoothing and generalized linear models to produce a more general model. The two techniques for estimation are introduced and illustrated.

In Section 5, we present a unified framework in which to view the estimation procedures. Section 6 contains examples of the procedures, including the logistic model and Cox's model for censored data. In Section 7 we discuss multiple covariate models and backfitting procedures.

Section 8 compares the *Local Scoring* and *Local Likelihood* procedures, Section 9 deals with their asymptotic properties, and in Section 10 we discuss inference for the models. In Section 11 we analyze the air pollution data which Breiman and Friedman (1982) analyzed using their ACE model.

Section 12 details estimation of the link function as well the the covariate functions, and shows the connection to the PACE model. Finally, in Section 13, we discuss the relationship of Generalized Additive Models to other models suggested in the literature.

2. The Linear Regression Model and its Smooth Extension.

Our discussion will center on a response random variable Y, and a set of predictor random variables $X_1, X_2, \cdots X_p$. A set of n independent realizations of these random variables will be denoted by $(y_1, x_{11}, \ldots x_{1p}), \ldots (y_n, x_{n1} \ldots x_{np})$. When working with a single predictor ($p = 1$), we'll denote it by X and its realizations by $x_1, x_2, \ldots x_n$.

A regression procedure can be viewed as a method for estimating $\mathbf{E}(Y \mid X_1, X_2, \ldots X_p)$. The standard linear regression model assumes a simple form for this conditional expectation:

$$\mathbf{E}(Y \mid X_1, X_2, \ldots X_p) = \beta_0 + X_1\beta_1 + \ldots X_p\beta_p. \tag{1}$$

Given a sample, estimates of $\beta_0, \beta_1, \ldots \beta_p$ are usually obtained by least squares.

The additive smooth model generalizes the linear regression model. In place of (1), we assume

$$\mathbf{E}(Y \mid X_1, X_2 \ldots X_p) = s_0 + \sum_{j=1}^{p} s_j(X_j) \tag{2}$$

where the $s_j(\cdot)$'s are smooth functions standardized so that $\mathbf{E}\, s_j(X_j) = 0$. These functions are estimated one at a time, in a forward stepwise manner. Estimation of each $s_j(\cdot)$ is achieved through a *scatterplot smoother*.

2.1. Scatterplot Smoothers.

Let's look first at the case of a single predictor. Our model is

$$\mathbf{E}(Y \mid X) = s(X) \tag{3}$$

(If there is only one smooth function, we suppress the constant term s_0 and absorb it into the function). To estimate $s(x)$ from data, we can use any reasonable estimate of $\mathbf{E}(Y \mid X = x)$. One class of estimates are the ***local average estimates***:

$$\hat{s}(x_i) = \mathbf{Ave}_{j \in N_i}[y_j] \tag{4}$$

where "**Ave**" represents some averaging operator like the mean and N_i is a *neighbourhood* of x_i (a set of indices of points whose x values are *close* to x_i). The only type of neighbourhoods we'll consider in this paper are *symmetric nearest neighbourhoods*. Associated with a neighbourhood is the *span* or *window* size w; this is the proportion of the total points contained in each neighbourhood. Assuming that the data points are sorted by increasing x value, a span w symmetric neighbourhood at x_i will contain $[wn]$ points; half to the left of x_i and half to the right. A formal definition is:

$$N_i = \{\max(i - \frac{[wn]-1}{2}, 1), ..., i-1, i, i+1, ... \min(i + \frac{[wn]-1}{2}, n)\} \tag{5}$$

We see that the neighbourhoods are truncated near the end points if $\frac{[wn]-1}{2}$ points are not available. The span w controls the smoothness of the resulting estimate, and must be chosen in some way from the data.

If **Ave** stands for arithmetic mean, then $\hat{s}(\cdot)$ is the *running mean*, the simplest possible scatterplot smoother. The running mean is not a satisfactory smoother because it creates large biasses at the endpoints and doesn't reproduce straight lines (i.e. if the data lie exactly along a straight line, the smooth of the data will not be a straight line). A slight refinement of the running average, *the running lines smoother* alleviates these problems. The running lines estimate is defined by

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_i \tag{6}$$

where $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ are the least squares estimates for the data points in N_i:

$$\begin{aligned} \hat{\beta}_{1i} &= \frac{\sum_{j \in N_i}(x_j - \bar{x}_i) y_j}{\sum_{j \in N_i}(x_j - \bar{x}_i)^2} \\ \hat{\beta}_{0i} &= \bar{y}_i - \hat{\beta}_{1i} \bar{x}_i \end{aligned} \tag{7}$$

and $\bar{x}_i = \frac{1}{n}\sum_{j\in N_i} x_j$, $\bar{y}_i = \frac{1}{n}\sum_{j\in N_i} y_j$.

The running lines smoother is the most obvious generalization of the least squares line. When every neighbourhood contains 100% of the data points, the smooth agrees exactly with the least squares line. Although very simple in nature, the running lines smoother produces reasonable results and has the advantage that the estimate in a neighbourhood can be found by updating the estimate of the previous neighbourhood. As a result, a running lines smoother can be implemented in an $O(n)$ algorithm, a fact that will become important when we use it as a primitive in other procedures. For the rest of this paper, a " **Smooth**$(\cdot)$" operation will refer to a running lines smoother for some fixed span.

It is important to note, however, that the running lines smooth plays no special role in the algorithms that are described in this paper. Other estimates of $\mathbf{E}(Y \mid X)$ could be used, such as a kernel or spline smoother. Except for the increased computational cost, these smoothers could be expected to work as well or better than the running lines smooth.

Finally, using **Smooth** as a building block, the full model (2) can be estimated in a forward stepwise manner. This is discussed in Section 7.

2.2. Span Selection and the Bias-Variance Tradeoff.

The running lines smoother requires a choice of span size w. Let's look at the extreme choices first. When $w = 0$, $\hat{s}(x_i)$ is just y_i. This is not a good estimate because it has a high variance and is not smooth. If $w = 2.0$ (that is every neighbourhood contains all the data points), * $\hat{s}(\cdot)$ is the global least squares regression line. This estimate is *too* smooth and will not pick up curvature in the underlying function, i.e. will be biassed. Hence the span size should be chosen between 0 and 2 to trade-off the bias and variability of the estimate.

A data-based criterion can be derived for this purpose if we consider the estimates of $\mathbf{E}(Y \mid X)$ as empirical minimizers of the (integrated) prediction squared error

$$PSE = \mathbf{E}(Y - s(X))^2$$

or equivalently the integrated mean squared error

$$MSE = \mathbf{E}(\mathbf{E}(Y \mid X) - s(X))^2.$$

Let $\hat{s}_w^{-i}(x_i)$ be the running lines smooth of span w, at x_i, having removed (x_i, y_i) from the sample. Then the ***cross-validation*** sum of squares is defined by $CVSS(w) = (1/n)\sum_1^n (y_i - \hat{s}_w^{-i}(x_i))^2$. One can show that $\mathbf{E}(CVSS(w)) \approx PSE$, using the fact that $\hat{s}_w^i(x_i)$ is independent of y_i. Thus it is reasonable to choose the span w that produces the smallest value of $CVSS(w)$. This criterion effectively weighs bias and variance based on the sample. Cross-validation for span selection is discussed in Friedman and Stuetzle (1982). Note that if we used the observed

* A neighbourhood with span 1.0 would only contain half the data at the endpoints

residual error $RSS = \sum_1^n (y_i - \hat{s}_w(x_i))^2$ to choose w, ($\hat{s}_w(x_i)$ being the fit at x_i with span w) we would get $w = 0$ and hence $\hat{s}(x_i) = y_i$. Not surprisingly, $\mathbf{E}(RSS) \neq PSE$. The point is that by choosing the span to minimize an estimate of *expected* squared error, we get a sensible estimate.

3. A Review of Generalized Linear Models (GLMs).

Generalized linear models (Nelder and Wedderburn, 1972) consist of a *random component*, a *systematic component*, and a *link function*, linking the two components. The response Y is assumed to have exponential family density

$$f_Y(y;\theta;\phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\} \tag{8}$$

where θ is the natural parameter, and ϕ is the scale parameter. This is the random component of the model. We also assume that the expectation of Y, denoted by μ, is related to the set of covariates $X_1, X_2 \ldots X_p$ by $g(\mu) = \eta$ where $\eta = \beta_0 + \beta_1 X_1 \ldots \beta_p X_p$. η is the systematic component and g is the link function. Note that the mean μ is related to the natural parameter θ by $\mu = b'(\theta)$; also, the most commonly used link for a given f is called the *canonical link*, for which $\eta = \theta$. It is customary, however, to define the model in terms of μ and $\eta = g(\mu)$ and thus θ does not play a role. Hence, when convenient we'll write $f_Y(y,\theta,\phi)$ as $f_Y(y,\mu,\phi)$.

Estimation of μ doesn't involve the scale parameter ϕ, so for simplicity this will be assumed known.

Given specific choices for the random and systematic components, a link function, and a set of n observations, the maximum likelihood estimate of $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1 \ldots \hat{\beta}_p)$ can be found by a Fisher scoring procedure. GLIM uses an equivalent algorithm called *adjusted dependent variable regression* . Given $\hat{\eta}^0$, (a current estimate of the linear predictor), with corresponding fitted value $\hat{\mu}^0$, we form the "adjusted dependent variable"

$$z^0 = \hat{\eta}^0 + (y - \hat{\mu}^0)(\frac{d\eta}{d\mu})_0 \tag{9}$$

Define weights W^0 by

$$(W^0)^{-1} = (\frac{d\eta}{d\mu})_0^2 V^0 \tag{10}$$

where V^0 is the variance of Y at $\mu = \hat{\mu}^0$. The algorithm proceeds by regressing z^0 on $1, x_1, \ldots x_p$ with weights W^0 to obtain an estimate $\hat{\beta}$. Using $\hat{\beta}$, a new $\hat{\mu}$ and $\hat{\eta}$ are computed. A new z is computed and the process is repeated, until the changes in $\hat{\beta}$ are sufficiently small. Nelder and Wedderburn show that the adjusted dependent variable algorithm is equivalent to the Fisher Scoring procedure. It is attractive because no special optimization software is required, just a subroutine that computes weighted least squares estimates.

Deviations between the data and the estimated model are measured using the *Deviance*, defined by

$$\mathbf{Dev}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2[l(\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}})] \tag{11}$$

where $l(\boldsymbol{\mu}) = \sum \log f_Y(y_i, u_i, \phi)$ is the log likelihood. In the case of Gaussian errors the deviance is identical to residual sum of squares (RSS) and in the more general case it enjoys the Pythagorean properties of the RSS.

A comprehensive description of generalized linear models is given by McCullagh and Nelder (1983).

4. Smooth Extensions of Generalized Linear Models.

4.1. Specification of the Model.

The linear predictor $\eta = \beta_0 + X_1\beta_1 \ldots X_p\beta_p$ specifies that $X_1, X_2, \ldots X_p$ act in a linear fashion. A more general model is

$$\eta = s_0 + \sum_1^p s_j(X_j) \tag{12}$$

where $s_1(\cdot) \ldots s_p(\cdot)$ are smooth functions. These functions will not be given a parametric form but instead will be estimated in a non-parametric fashion.

4.2. Estimation of the model — Local Scoring.

We require an estimate of the $s_j(\cdot)$'s in (12). For the linear model $\eta = \beta_0 + X_1\beta_1 + \ldots X_p\beta_p$, the estimates were found by repeatedly regressing the adjusted dependent variable z on $1, X_1, \ldots X_p$. Since smoothing generalizes linear regression, in the smooth model $\eta = s(X)$, we can estimate $s(\cdot)$ by smoothing the adjusting dependent variable on X. We call this procedure *Local Scoring* because the Fisher scoring update is computed using a local estimate of the score. This sensible idea can be justified on firm grounds (see Section 5); we display in Figure 1 (Section 6) the results of local scoring smoothing, $\exp(\hat{s}(x))/(1 + \exp(\hat{s}(x)))$, along with the usual linear estimate, $\exp(\hat{\alpha} + \hat{\beta}x)/(1 + \exp(\hat{\alpha} + \hat{\beta}x))$, for some binary response data. This is one of the smooths from the analysis of Haberman's breast cancer data, discussed in detail in Sections 6 and 7.

The span at each iteration is found by cross-validation, as described in Section 2. Recall that $\mathbf{E}(CVSS(w)) \approx PSE$ for a scatterplot smoother; the derivation of this rests on the fact that the fitted value for y_i does not involve y_i, and is thus independent of y_i. In this setting, the response is the adjusted dependent variable z_i which is a function of y_i. The cross-validated fit for z_i is a function of z_j, $j \neq i$. Since z_j is a function of y_i from previous iterations, z_i is not independent of its cross-validated fit. However, if k_n is the number of points in the neighbourhood, then one can show that under reasonable conditions the dependence is only $O(1/k_n)$.

To obtain smoother estimates, we use a slight modification of this criterion. We choose a larger span than the cross-validatory choice if it produces less than a 1% increase in $CVSS(w)$.

For the full model (12) , the smooths can be estimated one at a time in an iterative fashion. This idea is discussed in detail in Section 7.

4.3. Estimation of the model — Local Likelihood Estimation.

Tibshirani (1982), Hastie (1983a) and Tibshirani (1984) discuss another method for estimating smooth covariate functions, called *Local Likelihood* estimation. For a single covariate, the usual (linear) procedure fits a line across the entire range of X, i.e. $\eta = \beta_0 + \beta_1 X$. To estimate the model $\eta = s(X)$, the local likelihood procedure generalizes this by assuming that *locally* $s(x)$ is linear, and fits a line in a neighbourhood around each X value. In the exponential family with canonical link, the local likelihood estimate of $s(x_i)$ is defined as

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_i \tag{13}$$

where $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ maximize the local likelihood:

$$\log L_i = \sum_{j \in N_i} \{[y_j \theta_{ij} - b(\theta_{ij})]/a(\phi) + c(y_j, \phi)\} \tag{14}$$

and $\theta_{ij} = \beta_{0i} + \beta_{1i} x_j$. The local likelihood smooth applied to the Haberman data is also shown in Figure 1, along with the local scoring smooth. They are very similar, a fact that seems to be a general phenomenen. We discuss the relationship between the two procedures in Section 8.

Local scoring and local likelihood estimation provide two methods for estimating the covariate functions of a generalized linear model. In the next section, we introduce a theoretical framework in which to view both of these techniques. Besides providing a justification for the methods, this framework also produces a general form of local scoring that can be used in any likelihood-based regression model.

5. Justification of the Smoothing Procedures.

5.1. The Expected Log-likelihood Criterion.

In Section 2 we discussed scatterplot smoothers as estimates of $\mathbf{E}(Y \mid X = x)$. There we saw that by choosing the span to minimize an estimate of *expected* squared error, (as opposed to residual sum of squares) we obtained a sensible estimate. In this section, we will use this idea in a likelihood setting, basing the estimation procedures on *expected* log likelihood.

Consider a likelihood based regression model with one covariate. We assume that the data pairs $(x_1, y_1), \ldots (x_n, y_n)$ are independent realizations of random variables X and Y. Assume

also that given $X = x$, Y has density

$$Y \mid X = x \sim h(y, \eta) \tag{15}$$

Since η is a function of x, we will sometimes write $\eta(x)$ for emphasis. Denote the corresponding log-likelihood for a single observation by $l(\eta, Y)$, or l for short. Now to estimate $\hat{\eta}(\cdot)$, we could simply maximize $\sum_1^n l(\eta(x_i), y_i)$ over $\eta(x_1), \eta(x_2), \ldots \eta(x_n)$. This is unsatisfactory, however, because it doesn't force the estimate to be smooth. In the logistic model, for example, it produces $\hat{\eta}(x_i) = +\infty$ if $y_i = 1$ and $-\infty$ if $y_i = 0$, and the estimated probabilities are just the observed y_i's. Looking back at the scatterplot smoothing discussion, we see that the remedy in the random variable case is to choose $\hat{\eta}(\cdot)$ to maximize the *expected* log-likelihood:

$$\hat{\eta}(\cdot) = \max^{-1}[\, \mathbf{E}(l(\eta(X), Y))] \tag{16}$$

the expectation being over the joint distribution of X and Y. Clearly, this is the direct generalization of mean squared error minimization in the Gaussian case. Mean-squared error generalizes to (expected) Kullback-Leibler distance in non-Gaussian models, and maximization of the expected log-likelihood is equivalent to minimization of the Kullback-Leibler distance.

The use of expected log likelihood has also been suggested by Brillinger (1979) and Owen (1983, unpublished manuscript). Note that for each $X = x$, this theoretical criterion is maximized by the true value $\eta(x)$; this is a key fact in deriving the asymptotic properties of maximum likelihood estimates. In what follows, we show that standard maximum likelihood estimation for generalized linear models, local scoring, and local likelihood estimation can all be viewed as methods for empirically maximizing the expected log likelihood.

5.2. Derivation of the Estimation Techniques via Expected Log-Likelihood.

One way to use (16) for estimation of $\eta(x)$ would be to assume a simple form for $\eta(x)$, like $\eta(x) = \beta_0 + \beta_1 x$. The expectation in (16) could then be replaced by its sample analogue, and the resultant expression maximized over β_0 and β_1. This is nothing more than standard maximum likelihood estimation.

Now suppose (as is the point of this paper) that we don't want to assume a parametric form for $\eta(x)$. Differentiating (16) with respect to η, we get

$$\mathbf{E}(\frac{dl}{d\eta} \mid x)_{\hat{\eta}(x)} = 0 \tag{17}$$

Given some initial estimate $\eta^0(x)$, a first order Taylor series expansion gives the improved estimate

$$\eta^1(x) = \eta^0(x) - \frac{\mathbf{E}(\frac{dl}{d\eta^0} \mid x)}{\mathbf{E}(\frac{d^2l}{d\eta^{0^2}} \mid x)} \tag{18}$$

or

$$\eta^1(x) = \mathbf{E}\left[\eta^0(x) - \frac{\frac{dl}{d\eta^0}}{\mathbf{E}(\frac{d^2l}{d\eta^{0^2}}\,|\,x)}\,|\,x\right] \tag{19}$$

This provides a recipe for estimating $\eta(\cdot)$ in practice. Starting with some initial estimate $\eta^0(x)$, a new estimate is obtained using formula (19), replacing the conditional expectations by scatterplot smooths. The data algorithm analogue is thus

$$\eta^1(x) = \mathbf{Smooth}\left[\eta^0(x) - \frac{\frac{dl}{d\eta^0}}{\mathbf{Smooth}[\frac{d^2l}{d\eta^{0^2}}]}\right] \tag{20}$$

Since the variance of each of the terms in the brackets is approximately $\propto \mathbf{E}(\frac{d^2l}{d\eta^2})$, the smooth should use weights $\propto \mathbf{Smooth}(\frac{d^2l}{d\eta^2})^{-1}$ for efficient estimation. The data algorithm consists of repeated iterations of (20), until convergence.

In the generalized linear model case, we can simplify (19) before replacing $\mathbf{E}(\cdot\,|\,x)$ by **Smooth**. We compute $\frac{dl}{d\eta} = (y-\mu)V^{-1}\frac{d\mu}{d\eta}$, $\frac{d^2l}{d\eta^2} = (y-\mu)\frac{d}{d\eta}[V^{-1}\frac{d\mu}{d\eta}] - (\frac{d\mu}{d\eta})^2V^{-1}$, and $\mathbf{E}(\frac{d^2l}{d\eta^2}\,|\,x) = -(\frac{d\mu}{d\eta})^2V^{-1}$. Hence the update simplifies to

$$\eta^1(x) = \mathbf{E}\left[\eta^0(x) + (Y-\mu^0)\frac{d\eta}{d\mu^0}\,|\,x\right] \tag{21}$$

The data analogue is

$$\eta^1(x) = \mathbf{Smooth}[\eta^0(x) + (y-\mu^0)\frac{d\eta}{d\mu^0}] \tag{22}$$

with weights $(\frac{d\mu}{d\eta^0})^2V^{-1}$. This is exactly a smooth of the adjusted dependent variable, suggested on intuitive grounds in Section 4.

Note that we chose the form (19) instead of (18). In the case of distributions, they are the same because conditional expectation is a projection operator. Most smooths are not projections and thus the two forms are not equivalent in the data case. We chose (19) because in the Gaussian case it simplifies to $\hat{\eta}(x) = \mathbf{Smooth}(y)$ without any iterations, whereas (18) would require iterations even in this simple case.

The local likelihood procedure can also be viewed as an empirical method of maximizing $\mathbf{E}l(\eta(X),Y)$. Instead of differentiating this expression (as above), we write $\mathbf{E}l(\eta(X),Y) = \mathbf{E}(\,\mathbf{E}(l(\eta(X),Y)\,|\,X=x))$. Hence it is sufficient to maximize $\mathbf{E}(l(\eta(X),Y)\,|\,X=x)$ for each x. The corresponding data recipe can be derived as follows. Consider estimating $\eta(x)$ at some point $x=x_i$. An estimate of $\mathbf{E}(l(\eta(X),Y)\,|\,X=x_i)$ is

$$\hat{\mathbf{E}}(l(\eta(X),Y)\,|\,X=x_i) = (1/k)\sum_{j\in N_i} l(\eta(x_j),y_j) \tag{23}$$

where $k = \#$ of data points in N_i. Assuming $\eta(x) \approx \beta_{0i} + \beta_{1i}x$ for points in N_i, (23) is then maximized over β_{0i} and β_{1i}. The resulting estimate, $\hat{\eta}(x_i) = \hat{\beta}_{0i} + \beta_{1i}x_i$, is the local likelihood estimate as defined in Section 4.

The algorithms described here can be used in any likelihood-based regression model. As a technical point, note that in the exponential family, we linked the additive predictor $\eta = \sum_1^p s_j(X_j)$ to the distribution of Y via $\eta = g(\mu)$. In some non-exponential family models, μ is a complicated function of the model parameters or may not exist at all., It would then be desirable to link η to some other parameter of the distribution. This is true in the Cox model (see the next section). In any case, there is no difficulty — however η is linked to the distribution of Y, the likelihood is some function of η and its derivatives are used in the updating formula.

To summarize so far, maximization of the expected log-likelihood has led to a general technique for estimating a smooth covariate function: the local scoring procedure. In the case of the exponential family likelihood this procedure corresponds to smoothing of the adjusted dependent variable. Standard (linear) maximum likelihood estimation and local likelihood estimation can also be viewed as as empirical maximizers of expected log-likelihood. Equivalently they can all be viewed as empirical minimizers of the expected Kullback-Leibler distance between the model and the estimate.

We have not addressed the problem of multiple covariates — this will be done in Section 7.

5.3. Span selection.

In the Gaussian or ordinary additive regression models we use the CVSS to guide us in selecting spans. CVSS is approximately unbiassed for the *Expected Prediction Squared Error (PSE)*, whereas the RSS is not and would lead us to pick spans of 0. In the exponential family, the *Deviance* is the analogue of RSS. It is a sample estimate of the expected Kullback-Leibler distance between a model and future observations. Just like the RSS it will be biassed for this quantity. For span selection, one can think of cross-validating the deviance in order to get an approximately unbiassed estimate for the Kullback-Leibler distance. This, however, is very expensive due to the non-linear nature of the estimation procedures. In ordinary additive regression, simple deletion formulae allow one to compute cross validated fits in linear time. In the case of generalized additive models, however, the entire estimation procedure has to be repeated n times, and so cross-validation is infeasible.

Instead we use cross-validation to select the span each time we compute a smooth, as outlined in Section 4.2. This is done in linear time, and since squared error is a first order approximation to the deviance, it can be thought of as an approximation to the cross-validated deviance.

6. Some Examples.

6.1. The Gaussian Model.

For this model, $\eta = \mu$, so (22) simplifies to $\eta^1(x) = \textbf{Smooth}[y]$, and the local scoring algorithm reduces to a running lines smooth of y on x.

The local likelihood procedure also gives the running lines smooth of y on x, since the local m.l.e is $\hat{\eta}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i}x_i$, $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ being the least squares estimates for the points in N_i. The Gaussian model is applied to a large meteorological data set in Section 11.

6.2. The Logistic Model.

A binomial response model assumes that the proportion of successes Y is such that $n(x)Y \mid x \sim Bin(n(x), p(x))$, where $Bin(n(x), p(x))$ refers to the binomial distribution with parameters $n(x)$ and $p(x)$. Often the data is binary in which case $n(x) \equiv 1$. The binomial distribution is a member of the exponential family with canonical link $g(p(x)) = \log \frac{p(x)}{1-p(x)} = \eta(x)$. In the *linear logistic model* we assume $\eta(x) = \beta_0 + \beta_1 x$, and the parameters are estimated by maximum likelihood using Fisher's scoring or equivalently by using adjusted dependent variable regression. The smooth extension of this model generalizes the link relation to $\log \frac{p(x)}{1-p(x)} = \eta(x)$. The local scoring step is

$$\eta^1(x) = \textbf{Smooth}[\eta^0(x) + \frac{y - p^0(x)}{p^0(x)(1-p^0(x))}] \tag{24}$$

with weights $n(x)p^0(x)(1-p^0(x))$. We now demonstrate the procedure on some real data.

A study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital concerned the survival of patients who had undergone surgery for breast cancer (Haberman, 1976). There are 306 observations on four variables.

$$y_i = \begin{cases} 1 & \text{if patient } i \text{ survived 5 years or longer;} \\ 0 & \text{otherwise.} \end{cases}$$

x_{i1} = age of patient i at time of operation

x_{i2} = year of operation i (minus 1900)

x_{i3} = number of positive axillary nodes detected in patient i

Figure 1 shows the response variable plotted against the covariate *age*. The solid non-linear function was estimated using the local scoring method. For a single covariate one could simply average the 0-1 response directly, with iterative weights $[\hat{p}(x_i)(1-\hat{p}(x_i))]^{-1}$. This estimates $\mathbf{E}(Y \mid x) = p(x)$, and is the dashed curve in the figure. It is identical to the function found using the *local likelihood* method fitting local constants to the logits. The local likelihood smooth fitting local straight lines (the more usual approach) is the dotted curve. They are all very similar, with the main differences occuring in the tails where bias effects play a role. We

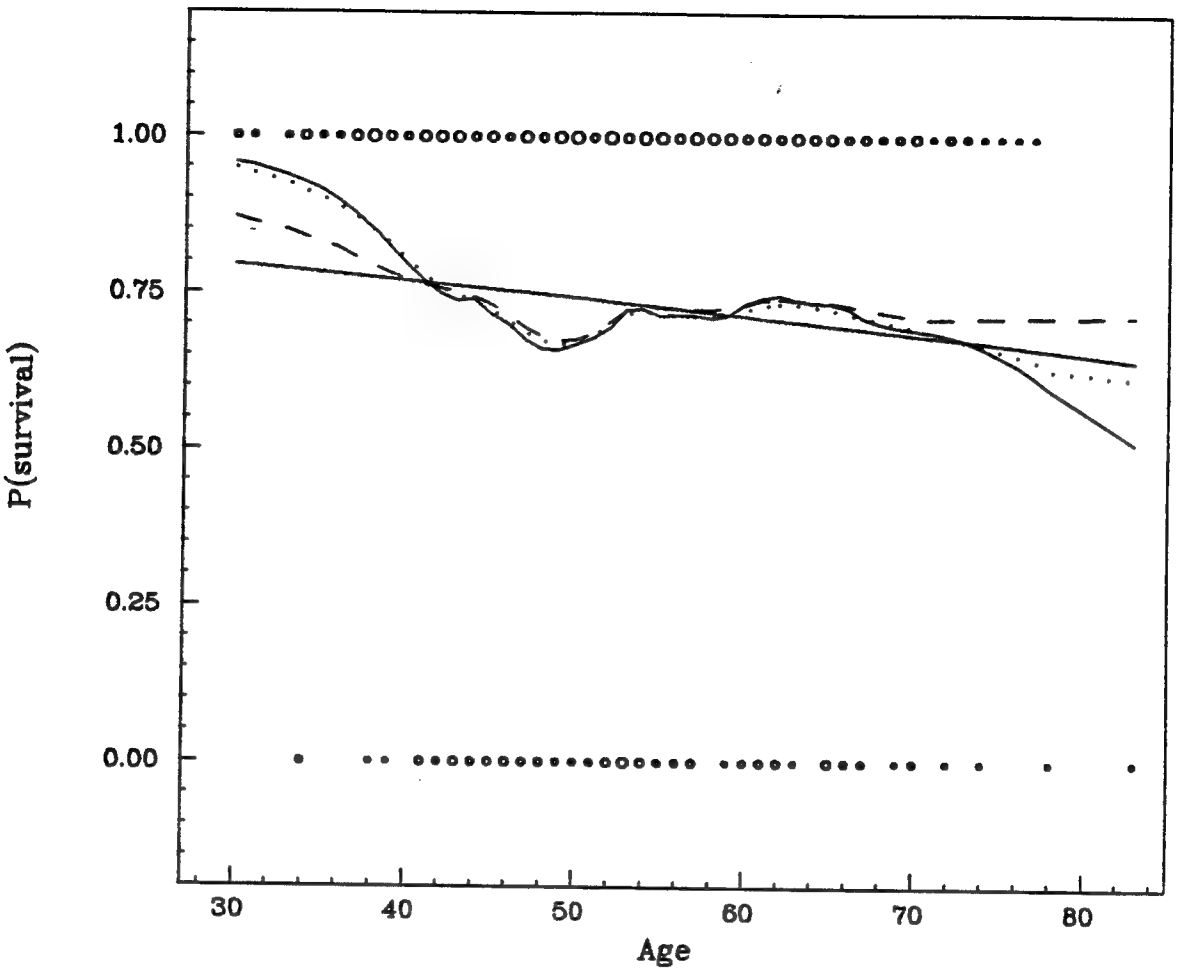


Figure 1. Survival of patients who underwent surgery versus age of the patient. The local scoring function is the solid curve, the local likelihood function is dotted, the running mean of the y's is dashed, and the linear logistic function is the almost straight curve. The size of the points reflects the number of observations.

will see in the Section 7 that in fitting multiple covariate models, this approach of smoothing the response variable directly breaks down, whereas the local scoring and local likelihood techniques generalize easily. We will pursue this example in Section 7.

6.3. The Cox Model.

The proportional hazards model of Cox (1972) is an example of a non-exponential family regression model. This model is used to relate a covariate to a possibly censored survival time. The data available are of the form $(y_1, x_1, \delta_1,) \ldots (y_n, x_n, \delta_n)$, the survival time y_i being

complete if $\delta_1 = 1$ and censored if $\delta_i = 0$. We assume there are no ties in the survival times. The proportional hazards model assumes the hazard relation

$$\lambda(t \mid x) = \lambda_0(t) e^{\beta x} \tag{25}$$

The parameter β can be estimated without specification of $\lambda_0(t)$ by choosing $\hat{\beta}$ to maximize the *partial likelihood*

$$PL = \prod_{i \in D} \frac{e^{\beta x_i}}{\sum_{j \in R_i} e^{\beta x_j}} \tag{26}$$

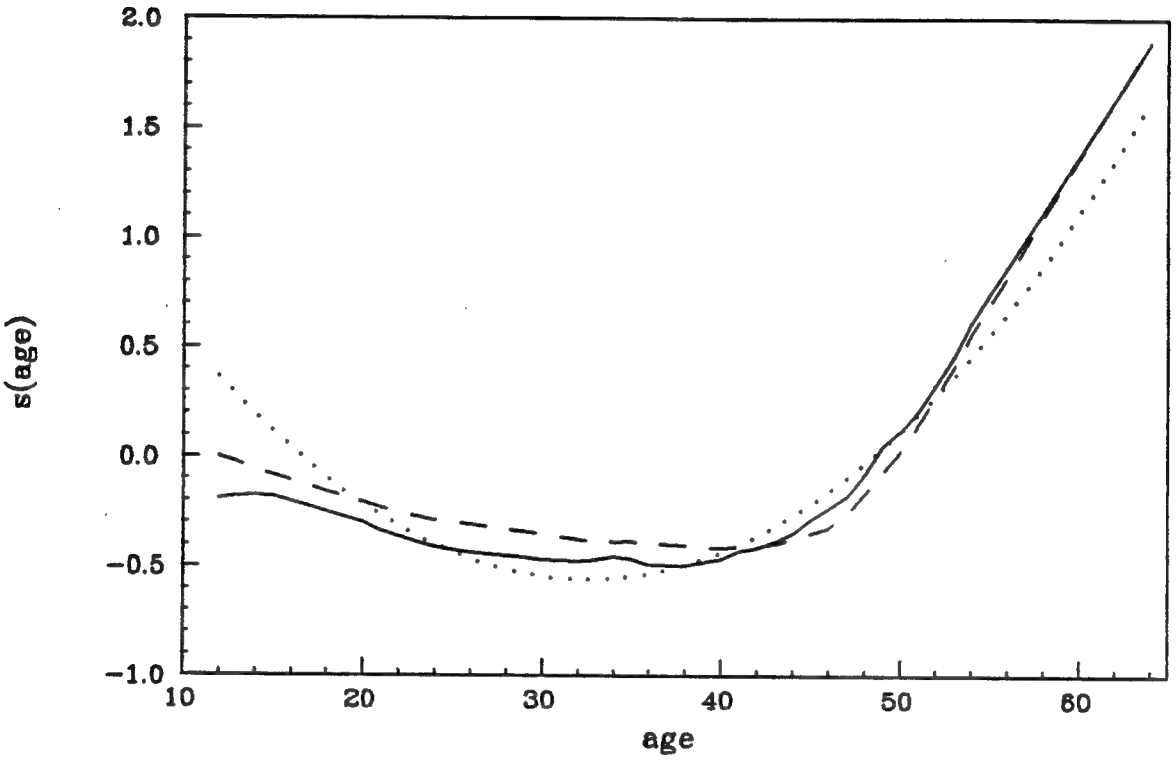


Figure 2. The Stanford Heart Transplant data. The solid curve is the local scoring function, the dashed line is the local likelihood function, and the dotted line is the parametric model.

In the above, D is the set of indices of the failures and R_i is the risk set just before the failure at y_i.

A more general model is

$$\lambda(t \mid x) = \lambda_0(t) e^{\eta(x)} \tag{27}$$

where $\eta(x)$ is a smooth function of x. One way to estimate $\eta(x)$ would be to apply the local scoring formula (20). Letting l equal the log-partial likelihood and $C_i = \{k : i \in R_k\}$, (the risk

sets containing individual i), straighforward calculations yield

$$\frac{\partial l}{\partial \eta(x_i)} = \delta_i - e^{\eta(x_i)} \sum_{k \in C_i} \frac{1}{\sum_{j \in R_k} e^{\eta(x_j)}} \tag{28}$$

and

$$\frac{\partial^2 l}{\partial \eta(x_i)^2} = -e^{\eta(x_i)} \sum_{k \in C_i} \frac{1}{\sum_{j \in R_k} e^{\eta(x_j)}} + e^{2\eta(x_i)} \sum_{k \in C_i} \frac{1}{(\sum_{j \in R_k} e^{\eta(x_j)})^2} \tag{29}$$

Starting with say $\eta(x) = \hat{\beta} x$, smooths are applied to these quantities, as in (20), and the process is iterated.

The local likelihood technique can also be applied to the Cox model — this is described in Tibshirani (1984). We won't give details here. Instead, we'll illustrate the two estimation techniques with a real data example.

Miller and Halpern (1983) provide a number of analyses of the Stanford heart transplant data. The data consist of time to failure (months) and two covariates, age (years) and T5 mismatch score. Here we will consider only the age variable.

Figure 2 shows the smooth obtained by local scoring (solid line) and local likelihood (broken line). Also shown is the fit obtained by inserting a linear and quadratic term for age (dotted line). The smooths are very similar and both show a marked non-linear effect. Table 1 summarizes the results of the smooth procedures as well as a standard (linear) Cox model.

Table 1. Stanford Heart Transplant Data

Analysis of Age

Model	-2 Log Likelihood	dof
Null	902.40	0
Linear	894.82	1
Linear + Quadratic	886.24	2
Local Likelihood (span .5)	884.65	2.95
Local Scoring (span .5)	884.66	2.95

The column labelled "dof" means degrees of freedom — this is explained in section 10. The smooths suggest that the log relative risk stays about constant up to age 45, then rises sharply. The quadratic model forces a parametric shape on the function, and misleadingly suggests that the relative risk drops then rises. The data set is analysed more thoroughly in Tibshirani (1984).

7. Multiple Covariates.

When we have p covariates, represented by the vector $\boldsymbol{X} = (X_1, X_2, \ldots X_p)$, a general model specifies $\mathbf{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \mu$ and $g(\mu) = \eta(\boldsymbol{x})$ where η is a function of p variables. We will first discuss the Gaussian case, and show why it is necessary to restrict attention to an additive model.

We assume

$$Y = \eta(\boldsymbol{X}) + \epsilon \tag{30}$$

where $\eta(\boldsymbol{x}) = \mathbf{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$, $Var(Y \mid \boldsymbol{x}) = \sigma^2$, and the errors ϵ are independent of $\boldsymbol{X}$. The goal is to estimate $\eta(\boldsymbol{x})$. If we use the least squares criterion $\mathbf{E}(Y - \eta(\boldsymbol{X}))^2$, the best choice for $\eta(\boldsymbol{x})$ is $\mathbf{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$. In the case of a single covariate, we estimated $\mathbf{E}(Y \mid X = x)$ by a scatterplot smooth, which in its crudest form is the average of those y_i in the sample for which x_i is close to x.

We could think of doing the same thing for multiple covariates: average the y_i for which $\boldsymbol{x}_i$ is close to $\boldsymbol{x}$. However, it is well known that smoothers break down in higher dimensions (Friedman and Stuetzle, 1981); the *curse of dimensionality* takes its toll. The variance of an estimate depends on the number of points in the neighbourhood. However, you have to look further for near neighbours in high dimensions, and consequently the estimate is no longer local and can be severely biassed. This is the chief motivation for the additive model, $\eta(\boldsymbol{x}) = s_0 + \sum_{j=1}^{p} s_j(x_j)$. Each function is estimated by smoothing on a single co-ordinate; we can thus include sufficient points in the neighbourhoods to keep the variance of the estimates down and yet remain local in each co-ordinate. Of course, the additive model itself may be a biassed estimate of the true regression surface, but hopefully this bias is much lower than that produced by high dimensional smoothers. The additive model is an obvious generalization of the standard linear model, and it allows easier interpretations of the contributions of each variable. In practice a mixture of the two will often be used:

$$\eta(\boldsymbol{x}) = s_0 + \sum_{j=1}^{q} s_j(x_j) + \sum_{j=q+1}^{p} \beta_j x_j. \tag{31}$$

In later sections we will discuss other models more general than the additive model.

7.1. Estimation — The Additive regression model.

We now turn to the estimation of $s_0, s_1(\cdot), \ldots, s_p(\cdot)$ in the additive regression model

$$\mathbf{E}(Y \mid \boldsymbol{x}) = s_0 + \sum_{j=1}^{p} s_j(x_j), \tag{32}$$

where s_0 is a constant and $\mathbf{E}\, s_j(X_j) = 0 \; \forall \; j$.

In order to motivate the algorithm, suppose the model $Y = s_0 + \sum_{j=1}^{p} s_j(x_j) + \epsilon$ is in fact correct, and assume we know $s_0, s_1(\cdot), \ldots, s_{j-1}(\cdot), s_{j+1}(\cdot), \ldots, s_p(\cdot)$. If we define the partial residual:

$$R_j = Y - s_0 - \sum_{k \neq j} s_k(X_k),$$

then $\mathbf{E}(R_j \mid x_j) = s_j(x_j)$ and minimizes $\mathbf{E}(Y - s_0 - \sum_{k=1}^{p} s_k(X_k))^2$. Of course we don't know the $s_k(\cdot)$'s, but this provides a way for estimating each $\hat{s}_j(\cdot)$ given estimates $\hat{s}_i(\cdot)$, $i \neq j$. The resulting iterative procedure is known as the ***backfitting*** algorithm (Friedman and Stuetzle, 1981):

Backfitting algorithm

Initialization: $s_0 = \mathbf{E}(Y)$, $s_1^0 \equiv s_2^0 \equiv \ldots \equiv s_p^0 \equiv 0$, $m = 0$.

Iterate: $m \leftarrow m + 1$

for $j = 1$ to p do:

$$R_j = Y - s_0 - \sum_{k=1}^{j-1} s_k^m(X_k) - \sum_{k=j+1}^{p} s_k^{m-1}(X_k).$$

$$s_j^m(x_j) = \mathbf{E}(R_j \mid x_j).$$

Until: $RSS = \mathbf{E}(Y - s_0 - \sum_{j=1}^{p} s_j^m(X_j))^2$ fails to decrease.

In the above s_j^m denotes the estimate of s_j at the mth iteration. Notice that by effectively centering Y at the start, we guarantee that $\mathbf{E}s_j^m(X_j) = 0$ at every stage. It is clear that RSS does not increase at any step of the algorithm, and therefore converges. Breiman and Friedman (1982, Theorem 5.19) show in the more general context of the ACE (Alternating Conditional Expectation) algorithm that the solution $\sum s_j^{\infty}(x_j)$ is unique and is therefore the best additive approximation to $\mathbf{E}(y \mid \boldsymbol{x})$. This does not mean that the individual functions are unique, since dependence amongst the covariates can lead to more than one representation for the same fitted surface. These results do not depend on the validity of either the additive model for $\mathbf{E}(Y \mid \boldsymbol{x})$ or the additive error assumption as in (30) .

If we return to the world of finite samples, we replace the conditional expectations in the backfitting algorithm by their estimates, the scatterplot smooths. Breiman and Friedman have proved:

- For a restrictive (impractical) class of smoothers, the algorithm converges.
- For a less restrictive class, the procedure is mean square consistent in a special sense. Suppose that the mth iteration estimate of s_j is $\hat{s}_j^m$, where the *hat* implies it is a function of the sample size n. Let s_j^m be the estimate of s_j at the mth iteration of the algorithm applied to the distributions. Then $\mathbf{E}(\hat{s}_j^m(X) - s_j^m(X))^2 \to 0$ as $n \to \infty$.

A special case arises if we use the least squares regression $\hat{a} + \hat{b}x_j$ to estimate $\mathbf{E}(\cdot \mid x_j)$ at every stage of the algorithm. We can once again invoke the Breiman and Friedman results for

this projection operator, which show that the algorithm converges to the usual least squares estimate of the *multiple* regression of Y on $\boldsymbol{x}$. This is true for both the usual data estimate or the estimate in distribution space as in Section 5. We give an elementary proof of this fact in Appendix A.

Although these results are encouraging, much work is yet to be done to investigate the properties of additive models. In multiple regression we need to worry about collinearity of covariates when interpreting regression coefficients; perhaps *cocurvity* has even worse implications when trying to interpret the individual functions in additive models. This would call for non-parametric analogues of linear principal components analysis — a standard device for determining lower dimensional *linear* manifolds in the data. Some work in this direction has been done (Hastie, 1983b, 1984b, Young et al., 1978).

If the purpose of our analysis is prediction, these problems are less important. We proceed in an exploratory spirit, and hopefully a sound bed of theory will develop around these as yet unanswered questions.

7.2. Backfitting in the Local Scoring Algorithm.

For multiple covariates the Local Scoring update (19) is given by

$$\eta^1(\boldsymbol{x}) = \mathbf{E}\left[\eta^0(\boldsymbol{x}) - \frac{\frac{\partial l}{\partial \eta^0}}{\mathbf{E}\left[\frac{\partial^2 l}{\partial \eta^{02}} \mid \boldsymbol{x}\right]} \mid \boldsymbol{x}\right] \tag{33}$$

and in exponential family case (21) is

$$\begin{aligned} \eta^1(\boldsymbol{x}) &= \mathbf{E}\left[\eta^0(\boldsymbol{x}) + (Y - \mu^0)\frac{\partial \eta}{\partial \mu^0} \mid \boldsymbol{x}\right] \\ &= \mathbf{E}(Z \mid \boldsymbol{x}) \end{aligned} \tag{34}$$

where $g(\mu^0) = \eta^0$ and Z is the adjusted dependent variable. For the reasons described in the previous section, we will restrict attention to an additive model:

$$\eta(\boldsymbol{x}) = s_0 + \sum_{j=1}^{p} s_j(x_j)$$

We see that (34) is of the same form as equation (32) , with Z playing the role of Y. Thus to estimate the s_j's, we fit an additive regression model to Z, treating it as the response variable Y in (32) . The sum of the fitted functions is η^0 of the next iteration. This is the motivation for the *generalized backfitting algorithm* which we give for the exponential family case as in (34) .

Generalized backfitting algorithm

Initialization: $s_0 = \mathbf{E}(Z)$, $s_1^0 \equiv s_2^0 \equiv \ldots \equiv s_p^0 \equiv 0$, $m = 0$.

Iterate: $m \leftarrow m + 1$

1) Form the adjusted dependent variable

$$Z = \eta^{m-1} + (Y - \mu^{m-1})\frac{\partial \eta}{\partial \mu^{m-1}},$$

where $\eta^{m-1} = s_0 + \sum_{j=1}^{p} s_j^{m-1}(X_j)$ and $\eta^{m-1} = g(\mu^{m-1})$.

2) Form the weights $W = (\frac{\partial \mu}{\partial \eta^{m-1}})^2 V^{-1}$.

3) Fit an additive model to Z using the backfitting algorithm with weights W, to get estimated functions s_j^m and model η^m.

Until: $D = \mathbf{E}\ \mathbf{Dev}(Y, \mu^m)$ fails to decrease.

Step 3 of the algorithm is simply the additive regression backfitting algorithm with weights. In Appendix B we show why weights are required even in the distribution version of the algorithm. To incorporate them, the data is first transformed using the weights, and the backfitting algorithm is then applied to the transformed data. From the results of the previous section, we see that the inner loop converges. In particular, if each smooth was replaced by the simple regression on the corresponding covariate (for data or distributions), the backfitting algorithm converges to the usual (weighted) multiple regression. This shows that in this case, the algorithm is identical to the usual GLM estimation procedure using Fisher scoring as in (9) and (10). Once again the data analogue of the algorithm replaces weighted conditional expectations by weighted smooths.

The backfitting idea is also used in the local likelihood estimation procedure to incorporate multiple covariates. To estimate a new s_j, or adjust s_j for other s_k in the model, s_j is re-estimated holding all others fixed. The algorithm cycles through the functions until convergence. The details can be found in Tibshirani (1984).

7.3. The breast cancer example continued.

We continue our analysis of the breast cancer data using all 3 covariates. The model is now $\log \frac{p(x)}{1-p(x)} = s_0 + \sum_{j=1}^{3} s_j(x_j)$. This is preferable to modelling $p(X)$ by an additive sum, since we would have to check that the estimated probabilities are positive and add to 1; the logit transform allows our estimates to be unrestricted. There are other reasons for using the logit transform; on the logit scale prior probabilities appear only as an additive constant (McCullagh and Nelder, 1983). This is useful in biomedical problems where there is often some established population risk, and the problem is to see what factors modify this risk for the sample under study.

Table 2 summarizes the various models fitted. The column labelled *dof* refers to the approximate degrees of freedom or ***number of parameters*** of the model. The derivation of these quantities is outlined in Section 10. ***Auto*** in the column labelled spans indicates that each time a smooth was computed, the span was selected by cross-validation. The entry D^2 refers to the percentage of deviance explained, and is in direct analogy to the more familiar R^2 in regression. Figures 3, 4, and 5 show the estimated functions for our model with deviance 308.22 and *dof* = 8.8.

Landwehr et al (1984) analysed this data set and in particular considered partial residual plots in order to identify the fundamental form in which terms should appear. Their final model was

$$\text{logit } p(\boldsymbol{x}) = \beta_0 + x_1\beta_1 + x_1^2\beta_2 + x_1^3\beta_3 + x_2\beta_4 + x_1x_2\beta_5 + (\log(1+x_3))\beta_6 \tag{35}$$

with a deviance of 302.3 on 299 *dof*. We fit this model using GLIM, and then using the backfitting procedure with linear fits for the transformed variables. As expected, the results agreed up to 4 significant figures, which provides an empirical proof of the result proved in Appendix A. This model is labelled ***parametric*** in the table. We have superimposed their parametric model terms in the figures, and note that the functions are very similar. If $\boldsymbol{x}_i^t\boldsymbol{b}$ is the estimated linear model, and p_i^l the corresponding probability estimate, the partial residual for variable j and observation i is defined by

$$r(x_{ij}) = b_j x_{ij} + \frac{y_i - p_i^l}{p_i^l(1-p_i^l)}. \tag{36}$$

Landwehr et al. show that if the true model is logit $p(\boldsymbol{x}) = \beta_0 + \sum_{k\neq j}\beta_k x_k + s_j(x_j)$, then $\mathbf{E}[r(X_j) \mid X_j = a] \approx s_j(a)$. They then use the smooth of the partial residuals to suggest the functional form. This result breaks down if the other terms are not linear (Hastie, 1984a and Gong, 1984). One can see from the previous section that smoothing the partial residual corresponds to the first step of the generalized backfitting procedure in the local scoring algorithm, if our starting guess is the linear model. The backfitting procedure continues, however, by simultaneously estimating and adjusting non-parametric functions for all the covariates.

8. Comparison of Local Scoring to Local Likelihood Estimation.

In a number of examples that we have tried, the *Local Scoring* and *Local Likelihood* procedures give very similar results. This is not surprising in light of the discussion of Section 5, where both techniques are viewed as empirical estimates of $\mathbf{E}(\log L)$. The difference seems to be in computational speed: local scoring is $O(n)$ while local likelihood, if the span increases like n^c, is $O(n^{c+1})$. For large data sets, the local scoring procedure is considerably faster. This leads us to ask: will the two procedures always give similar estimates? Artificially, they could be made very different. The reason for this is as follows. For a single covariate, the local likelihood

Table 2. The Analysis of Deviance (ANODEV) table for the breast cancer data.

Model	Spans	dof	Deviance	D^2
Constant	1		353.67	
x_1, x_2 & x_3	all linear	4	328.75	.07
x_1, x_2 & x_3	all .5	8 .8	307.89	.13
x_1, x_2 & x_3	auto	8 .0	308.22	.13
x_2 & x_3	auto	5 .9	317.66	.10
x_1 & x_3	auto	5 .0	312.68	.12
x_1 & x_2	auto	4 .1	346.71	.02
Parametric		7	302.30	.15

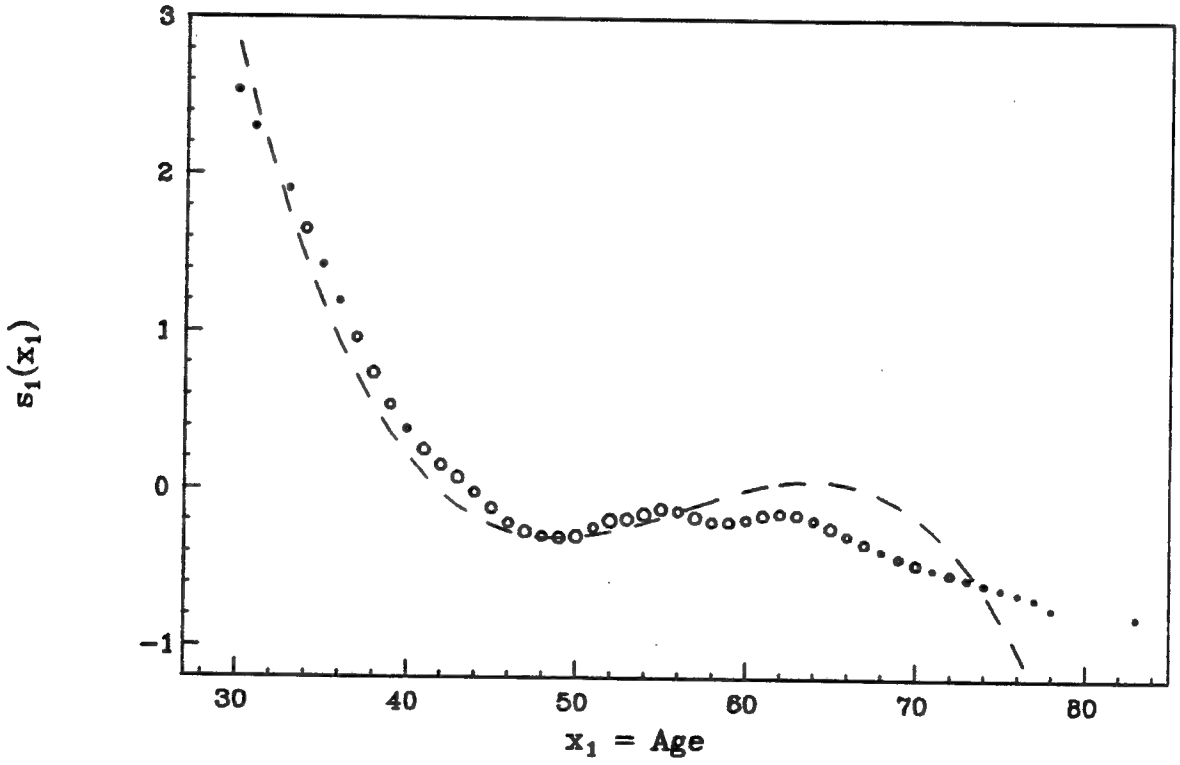


Figure 3. The circles represent $\hat{s}$(age), where the area of the circles is proportional to the number of points. The dashed term is the cubic polynomial term in (35)

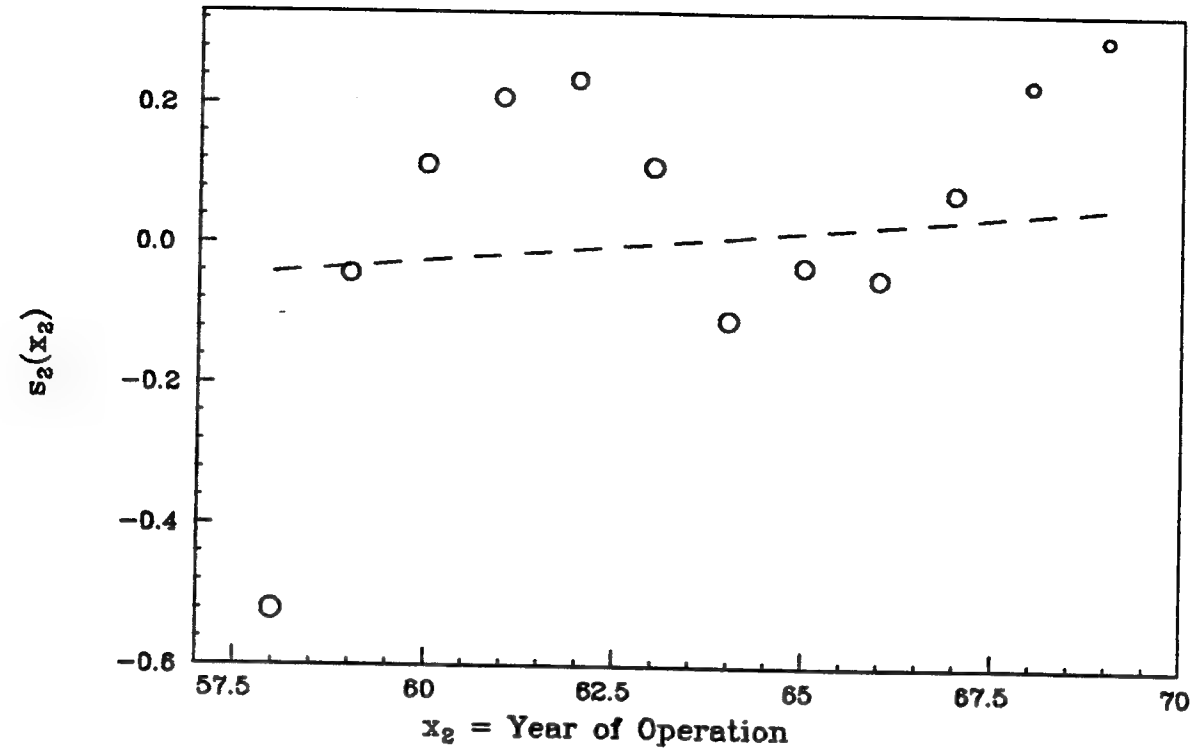


Figure 4. The circles represent $\hat{s}$(year of operation). The dashed term is the linear term in (35) .

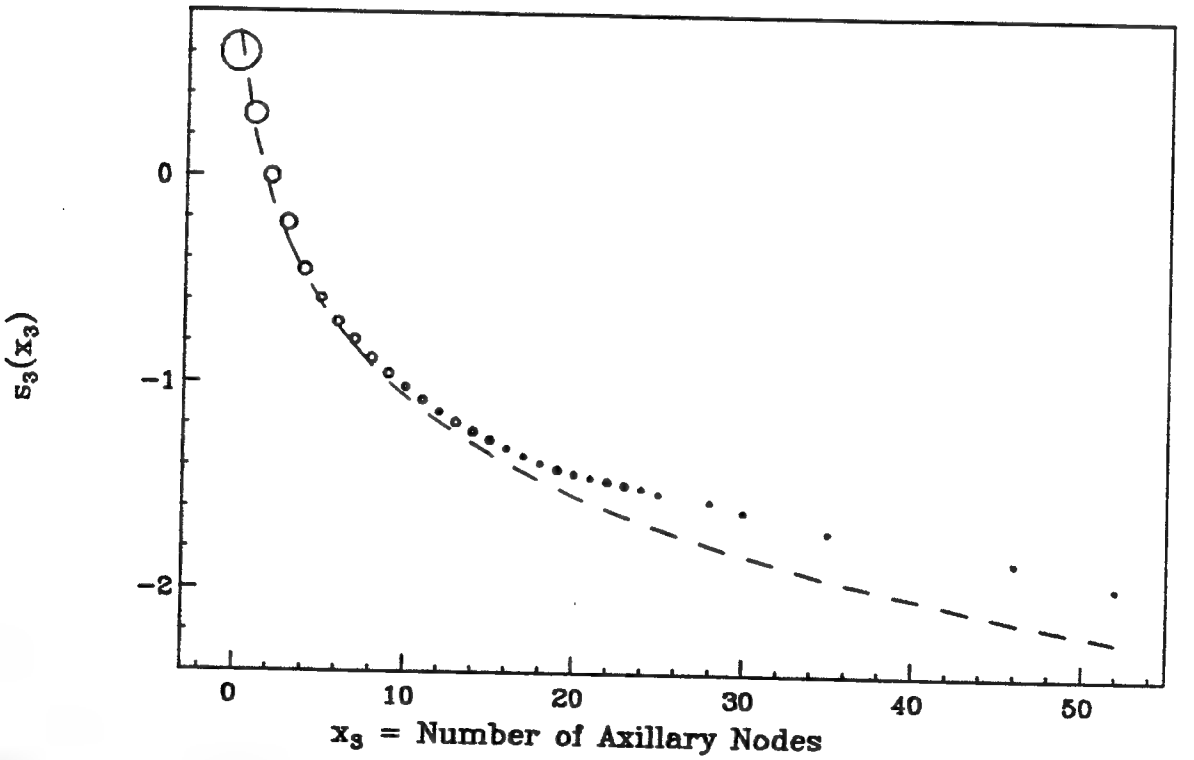


Figure 5. The circles represent $\hat{s}$(# of positive axillary nodes). The dashed term is the log term in (35) .

procedure is completely local; that is if x_j is not in the neighbourhood for estimating $\eta(x_i)$, then (x_j, y_j) has absolutely no effect on the estimate $\hat{\eta}(x_i)$. This is not true in the local scoring procedure, for as the smooth operation is iterated, the estimates $\hat{\eta}(x_j)$ enter into the computation of $\hat{\eta}(x_i)$. Thus sending y_j off to $+\infty$ would have a large effect on the estimate of $\hat{\eta}(x_i)$ in the smooth updating procedure, but no effect in the local likelihood procedure.

Given the theoretical basis of Section 5, it seems eminently reasonable that the two procedures be asymptotically equivalent. We sketch a proof of this fact in Appendix C.

For finite samples we can describe operationally the difference as follows, using logistic regression as an example. Suppose we start with $p^0(x_i) = \bar{y}$, the overall proportion of 1's. Then the first iteration for both procedures is identical:

- Local scoring fits the weighted least squares regression of $z_j = \text{logit } p^0 + \frac{y_j - p^0}{p^0(1-p^0)}$ against x_j for $j \in N_i$ to obtain the estimate $\eta^1(x_i)$; this is the local linear smoother operation in this neighbourhood.

- Local likelihood does exactly the same operation in computing the MLE in the neighbourhood, since this is the first step in the adjusted variable regression procedure used to compute the MLE.

The second iterations are very similar:

- Local scoring regresses $z_j = \eta^1(x_j) + \frac{y_j - p^1(x_j)}{p^1(x_j)(1-p^1(x_j))}$ against x_j for $j \in N_i$ to obtain the estimate $\eta^2(x_i)$.

- Local likelihood, however, regresses $z_j = \eta_i^1(x_j) + \frac{y_j - p_i^1(x_j)}{p_i^1(x_j)(1-p_i^1(x_j))}$ against x_j, where $\eta_i^1(x_j)$ refers to the extrapolated value of η^1 at x_j derived from the linear estimate $\eta^1(x_j) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_j$.

If the function is fairly linear in the neighbourhood then these two steps will yield similar estimates. For a given point x_j, the local scoring algorithm uses its latest estimate of $p(x_j)$ for every neighbourhood in which x_j appears. The local likelihood procedure, however, uses a linear approximation (on the η scale) for $p(x_j)$ based on its estimate $p(x_i)$ when x_j is in N_i.

9. Asymptotic Properties.

Since the local scoring and local likelihood procedures use local maximum likelihood estimates, we would expect them to have reasonable asymptotic properties. Tibshirani (1984) extends the work of McCullagh (1983) to establish such properties for local likelihood estimates in the exponential family. Consider estimation of a single smooth $\eta(\cdot)$ at a fixed point x_0. Let k_n be the number of points in the neighbourhood N_0^n used to estimate $\eta(x_0)$. Assume that $k_n \to \infty$, but the neighbourhood shrinks in such a way that $\max_{\{i,j \in N_0^n\}} |x_i - x_j| = o(k_n^{-1/2})$. Then under smoothness constraints on $\eta(\cdot)$, regularity conditions on the distribution of Y and boundedness on the covariate values, Tibshirani shows that the local likelihood estimate $\hat{\eta}(x_0)$

is consistent for the true value $\eta(x_0)$, and has the efficiency of a maximum likelihood estimator based on k_n (instead of n) observations. The proofs of these results rest on the fact that the local likelihood estimate of $\mathbf{E}(\log L \mid x_0)$ (see Section 5) is consistent. The results then follow by Taylor series expansions.

Similar results should be obtainable for the local scoring algorithm. They could be derived from the asymptotic equivalence with Local Likelihood estimation, or directly as follows. Consider the stable point of the update step (20). Assuming the neighbourhoods behave as above, $\mathbf{Smooth}[\eta^0(x)] \approx \eta^0(x)$ for large n, so at convergence $\mathbf{Smooth}[\frac{dl}{d\eta^0}] \approx 0$. Under conditions such that $\mathbf{Smooth}[\cdot]$ at x_0 is consistent for $\mathbf{E}(\cdot \mid x_0)$, the asymptotic properties of $\hat{\eta}(x_0)$ will follow from standard Taylor series arguments.

10. Deviance and Degrees of Freedom.

In generalized linear models, the goodness of fit of an estimate $\hat{\mu}$ is measured by the deviance. Wald's theorem tell us that, given two nested *linear* models and the hypothesis that the smaller model is correct, the deviance decrease in fitting the larger model is asymptotically $\sigma^2\chi^2_{p_2-p_1}$, where p_1 and p_2 are the ranks of the two linear spaces. That is, the number of parameters fit give the number of *degrees of freedom* of the corresponding deviance decrease.

This leads us to ask similar questions for the smooth estimates described in this paper. We will restrict our discussion to the exponential family case, and also to the case of known variance $\sigma^2 = 1$. The question of interest is: *how many "parameters" does a smooth employ?* This will depend on the span. With a span of 2.0 (i.e. every neighbourhood contains all the data points), 2 parameters are used. With a span of 0 (i.e. 1 point per neighbourhood), n independent parameters are used. Thus for the usual spans (.3 to .7), the number of parameters should be somewhere between 2 and n.

For the local likelihood procedure, Tibshirani(1984) provides a definition of "number of parameters" or "degrees of freedom" and an approximate method for determining it. This follows work by Cleveland (1979) on degrees of freedom for scatterplot smoothers.

Consider first a multiple linear regression model with variance equal to 1. Let $\hat{y}_1$ and $\hat{y}_2$ be the fitted values for two nested models, and assume that the sub-model, say model 1, is correct. Then the decrease in residual sum of squares has expected value $p_2 - p_1$. One way to derive this is to write $\hat{y}_1 = H_1 y$ and $\hat{y}_2 = H_2 y$, where H_1 and H_2 are the corresponding hat matrices. Then a simple calculation shows that the expected decrease in residual sum of squares is just $trace(H_2) - trace(H_1) = p_2 - p_1$.

For a running lines smoother, an analogous result can be derived. Consider a single covariate x. First, we note that the output of the smoother can be written as $\hat{y} = Sy$ where S is a "smoother matrix". Now consider two fit vectors $\hat{y}_1$ and $\hat{y}_2$ obtained by smoothing y on x with different spans. By analogy to the multiple linear regression case, we can define the difference in degrees of freedom of the fits by the expected value of $RSS(y, \hat{y}_2) - RSS(y, \hat{y}_1)$. In the regression set-up, this expected value was computed under the assumption that the

smaller model was correct. Here we make the analogous assumption that $\mathbf{E}\hat{y}_1 \approx \mathbf{E}\hat{y}_2$, i.e. that the two fits are about the same on the average. Then under this assumption, it is easy to show that $\mathbf{E}(RSS(y,\hat{y}_2) - RSS(y,\hat{y}_1)) = trace(S_2) - trace(S_1)$. Hence we can think of $trace(S)$ as the number of degrees of freedom of the smoother based on S.

Further, the same result is approximately true for any local likelihood fit in the exponential family. Consider two fits $\hat{\mu}_1$ and $\hat{\mu}_2$ based on a covariate vector x and spans l_1 and l_2. Then the difference in degrees of freedom of the two smooths is defined to be $\mathbf{E}(\mathbf{Dev}(\hat{\mu}_1, y) - \mathbf{Dev}(\hat{\mu}_2, y))$ under the assumption that $\mathbf{E}\hat{\mu_1} \approx \mathbf{E}\hat{\mu_2}$. Let S_1 and S_2 be the smoother matrices, based on x that produce running lines smooths of spans l_1 and l_2 respectively. Then it can be shown that $\mathbf{E}(\mathbf{Dev}(\hat{\mu}_1, y) - \mathbf{Dev}(\hat{\mu}_2, y)) \approx trace(S_2) - trace(S_1)$.

Given a local likelihood fit, we easily work out $trace(S)$ and use it to determine the significance of the smooth. As a example, for 200 equally spaced X values and a span of .5, $trace(S)$ is about 3.6. Hence the smooth uses roughly 3.6 parameters. We say "roughly" because the distribution of this decrease is not χ^2, however, but is more spread out. The $trace(S)$ formula should only be used as a rule of thumb.

Since the local scoring procedure produces estimates similar to local likelihood estimates, the same result should approximately be true for it as well. Due to the complex nature of the estimation procedure, however, we have been unable to verify this analytically. Instead, we describe a small simulation study designed to check the result numerically.

10.1. Degrees of Freedom Simulations.

Table 3. Results of Degrees of Freedom Simulation. Entries in Lines (2)—(7) are mean(variance) of deviance decrease. LS and LL mean local scoring and local likelihood respectively.

Source	Span .3	.4	.5	.6	.7
(1) Trace(S)-1	4.09	3.32	2.65	2.34	2.16
(2) Scatterplot Smooth(y normal)	4.14(10.00)	3.39(7.75)	2.61(6.03)	2.31(5.08)	2.09(4.32)
(3) Scatterplot Smooth(y uniform)	4.19(10.06)	3.46(8.50)	2.77(6.52)	2.41(5.79)	2.21(4.99)
(4) Logistic Model (LS) (constant vs smooth)	4.35(13.86)	3.39(11.78)	2.67(9.16)	2.35(8.02)	2.25(6.50)
(5) Logistic Model (LS) (linear vs smooth)	3.31(9.54)	2.36(8.27)	1.61(5.12)	1.33(4.12)	1.21(3.66)
(6) Logistic Model (LL) (constant vs smooth)	4.34(13.47)	3.40(11.62)	2.72(9.12)	2.28(7.51)	2.17(6.28)
(7) Logistic Model (LL) (linear vs smooth)	3.29(11.71)	2.25(8.25)	1.63(6.21)	1.29(4.58)	1.12(2.89)

Table 3 shows the results of a modest simulation study designed to check the accuracy of the formula $\mathbf{E}(D(y,\hat{y}_1) - D(y,\hat{y}_2)) = trace(S_2) - trace(S_1)$. The numbers in the table were obtained as follows. 100 x values were generated from $\mathcal{N}(0,1)$ and fixed for the entire table.

Given these x values, we constructed the running lines smoother matrices for the indicated spans, and the trace of each matrix (minus 1) is shown in line (1).

Consider for example the entry 4.09 in the top left hand corner. According to the discussion of the preceding section, this should be the expected decrease in deviance due to fitting a local likelihood or scoring model with that span .3 versus a model with only a constant.

To obtain line (2), we generated 100 y_i's from a $\mathcal{N}(0,1)$ distribution and computed $RSS(\boldsymbol{y},\bar{y}1) - RSS(\boldsymbol{y},\hat{\boldsymbol{y}})$, $\hat{\boldsymbol{y}}$ being the fit from a scatterplot smoother ($\hat{\boldsymbol{y}} = P\boldsymbol{y}$) with span as shown. Line(2) shows the mean and variance from 500 such repetitions of this process.

Line (3) was obtained in same way as line (2), except that the y_i's were generated from uniform $(-\sqrt{3},\sqrt{3})$, the range chosen so that $Var(y_i) = 1$.

To obtain line(4), we generated 100 y_i's from $binomial(1,1/2)$ and fit a local scoring logistic model with spans of .3 to .7. The numbers show the mean and variance of $D(\boldsymbol{y},\bar{y}1) - D(\boldsymbol{y},\hat{\boldsymbol{y}})$ over 500 repetitions.

Line (5) was generated in a similar fashion as line (4), showing instead the mean and variance of $D(\boldsymbol{y},\hat{\boldsymbol{y}}_l) - D(\boldsymbol{y},\hat{\boldsymbol{y}})$, $\hat{\boldsymbol{y}}_l$ being the linear logistic fit, with y_i generated from a linear logistic model, $P(y_i = 1 \mid x) = e^{.2x}/(1+e^{.2x})$.

Lines (6) and (7) are the same as (4) and (5) except that the smooths were obtained by local likelihood estimation.

Note that for all the models, a span of 2.0 gives either exactly or asymptotically a mean value of 1 (by Wald's theorem), and $trace(S) - 1$ is also equal to 1.

The results give fairly strong support to the approximation $\mathbf{E}(D(\boldsymbol{y},\hat{\boldsymbol{y}}_1) - D(\boldsymbol{y},\hat{\boldsymbol{y}}_2)) = trace(S_2) - trace(S_1)$. Lines (2) and (3) agree well with (1), not surprising since the approximation is exact for scatterplot smoothers. Line (4) also is in good agreement, with a small upward bias for smaller spans. Line (5) should be 1 less than line (1), (since the global linear fit uses 2 degrees of freedom) and the results indicate that. As we expected, the local likelihood results are very similar to the local scoring numbers.

The variance results are a little unsettling. The variance to mean ratio is often greater than 2 (the ratio for a chi-square variate), especially for the non-Gaussian models.

We conclude from these simulations that the approximation $\mathbf{E}(D(\boldsymbol{y},\hat{\boldsymbol{y}}_1) - D(\boldsymbol{y},\hat{\boldsymbol{y}}_2)) = trace(S_2) - trace(S_1)$ is satisfactory as a rough rule of thumb, for the Gaussian and logistic models. We do note, however, that the distribution of this decrease is more spread out than a chi-square variate with the corresponding degrees of freedom, so that tests based on the percentile points will be too liberal.

The numbers reported here for local likelihood estimation can also be found in Tibshirani(1984). In that study, the Cox model was also included in the simulation, and the *trace* formula was found to be biassed downward. Thus to accurately assess the significance of a Cox smooth in real examples, the mean decrease must be found by simulation. This was done for the example of Section 6.

11. Example: Ozone Concentration data.

In this section, we apply the local scoring procedure to some data on atmospheric ozone concentration, given in Breiman and Friedman (1982). The data consist of 330 observations on 10 variables:

Response:

UPO3: Upland Ozone Concentration (ppm)

Predictors:

VDHT: Vandenburg 500 millibar height (m)

HMDT: Humidity (percent)

IBTP: inversion Base Temperature (F^o)

SBTP: Sandburg Air Force Base Temperature (C^o)

IBHT Inversion Base Height (feet)

WDSP: Wind Speed (mph)

DGPG: Daggot Pressure Gradient (mmhg)

VSTY: Visibility (miles)

DOYR: Day of Year

The objective is to study the effect of various meterological variables on atmospheric ozone concentration. Following Breiman and Friedman, we considered all the covariates except DOYR initially, then examined the effect of adding DOYR to our model.

We used the Gaussian additive model to explore these data. A summary of the models is given in Table 4.

Table 4. The ANOVA table for the Ozone Concentration Data.

Model	Spans	dof	Deviance (RSS)	R^2
Constant	1		21115.41	
First 8 predictors	all linear	9	6539.00	.69
First 8 predictors	all .5	22 .5	5176.56	.75
All 9 predictors	auto	21 .8	4292.28	.80
SBTP, IBTH, DGPG, VSTY	auto	11 .0	5431.93	.74
SBTP, IBTH, DGPG, VSTY, DOYR	auto	12 .4	4736.60	.78
Semi-parametric		11 .2	4848.99	.77

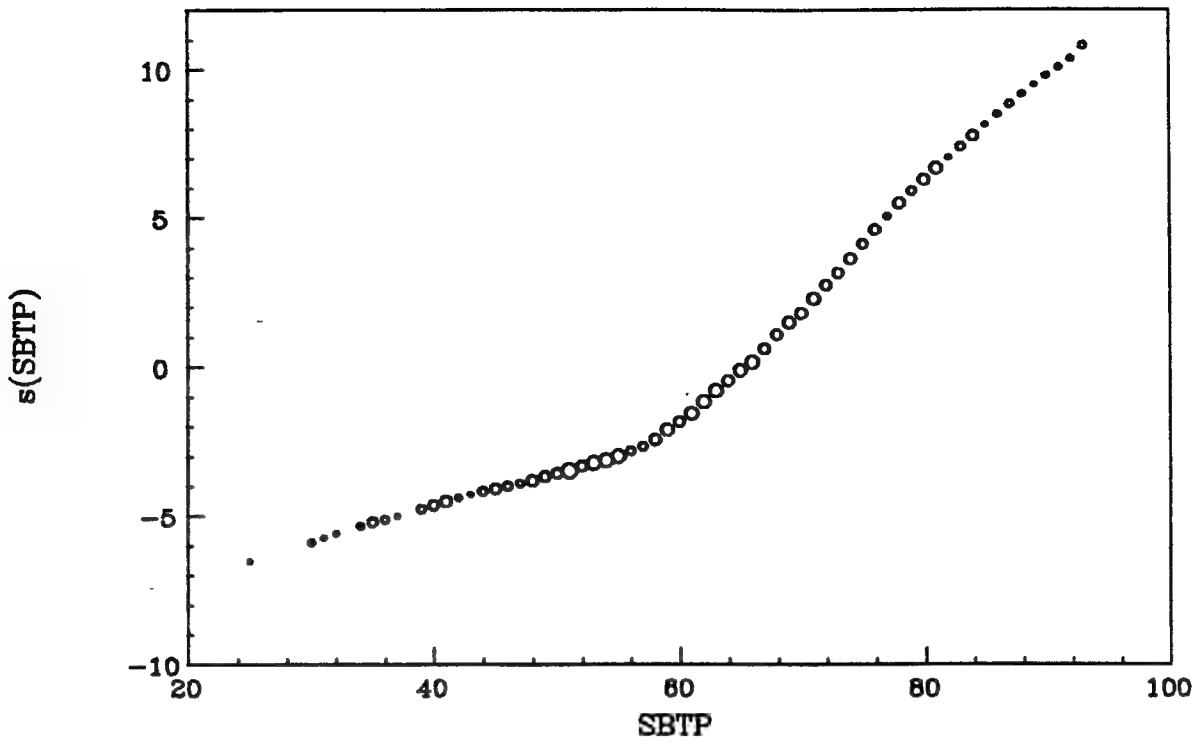


Figure 6. The estimated function for *Sandburg Air Force Base Temperature*. The area of the circles is proportional to the number of data points occurring at that value of SBTP.

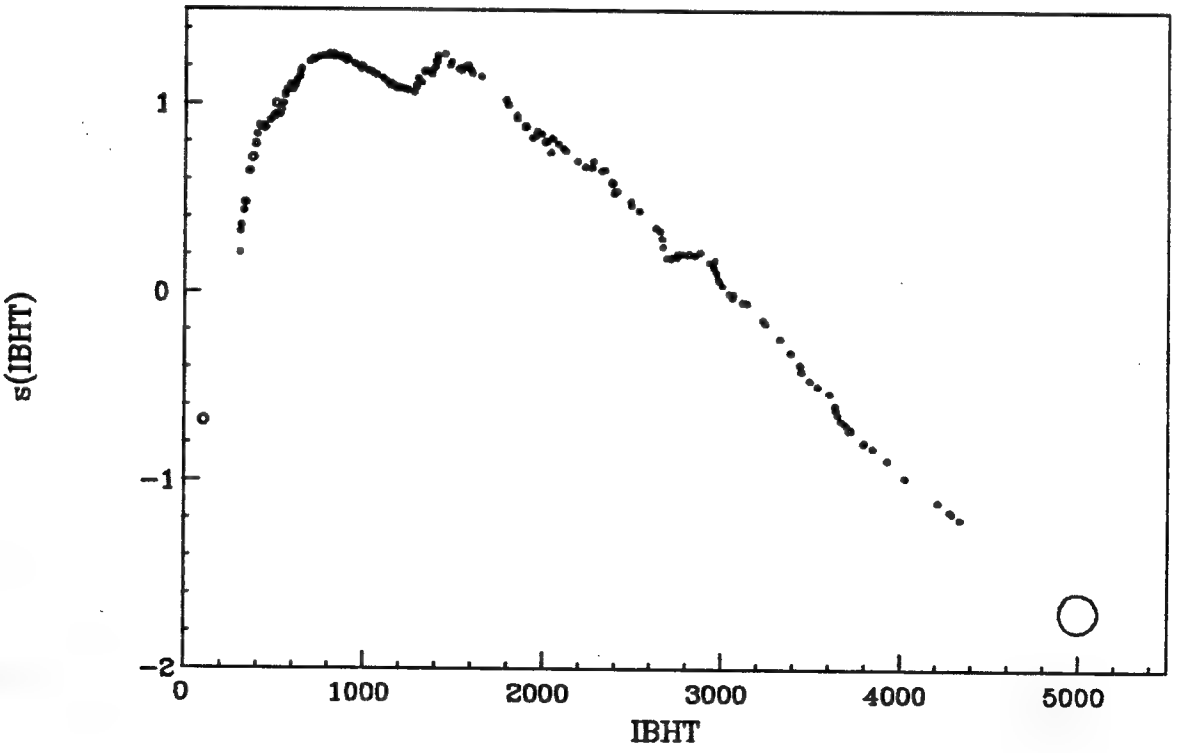


Figure 7. The estimated function for *Inversion Base Height*.

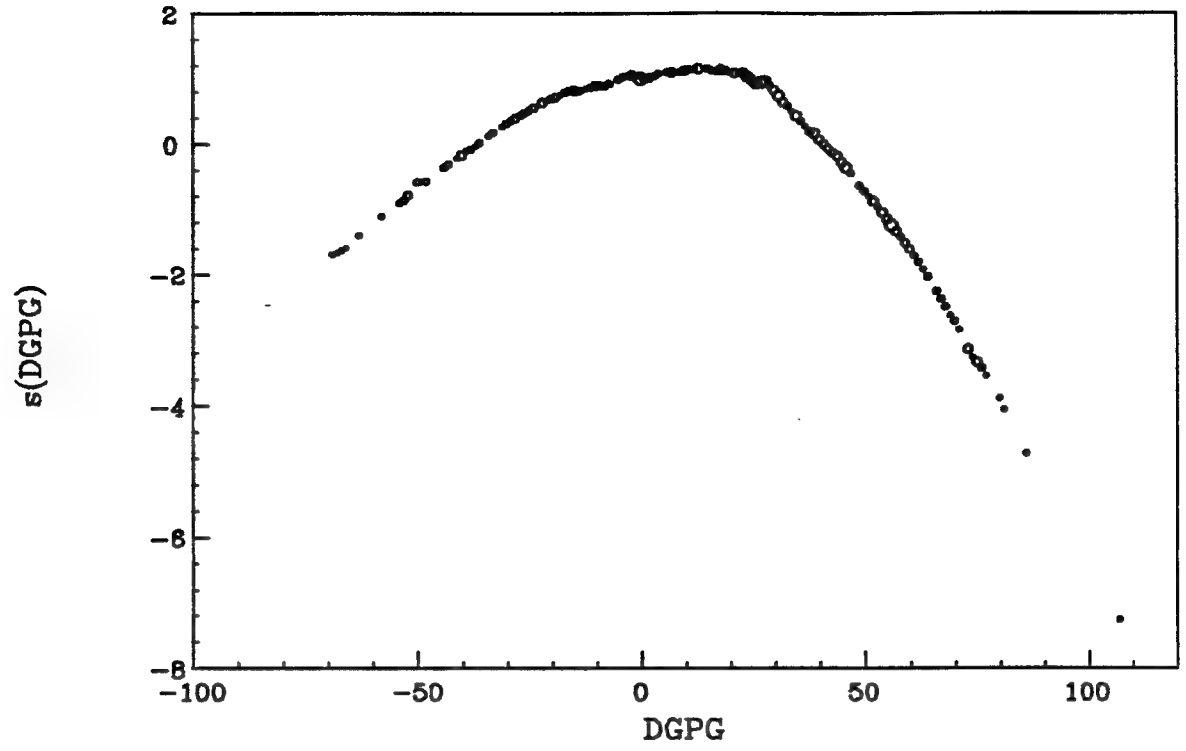


Figure 8. The estimated function for *Daggot Pressure Gradient.*

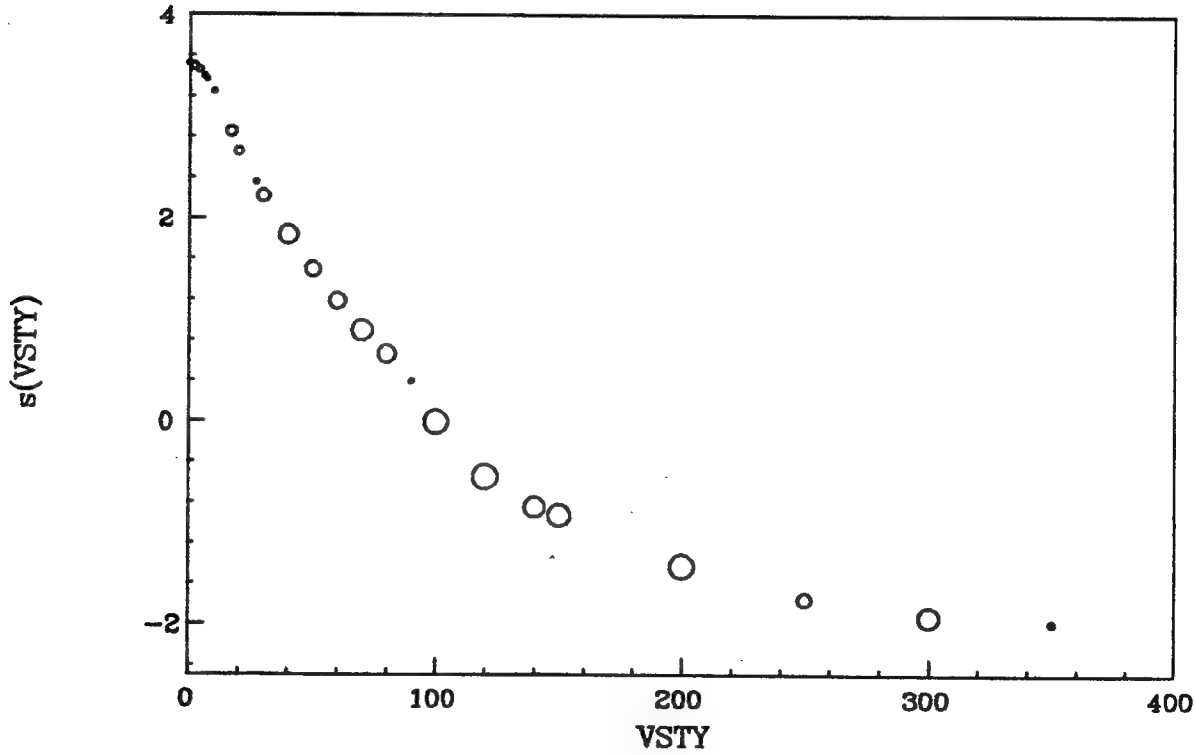


Figure 9. The estimated function for *Visibility.*

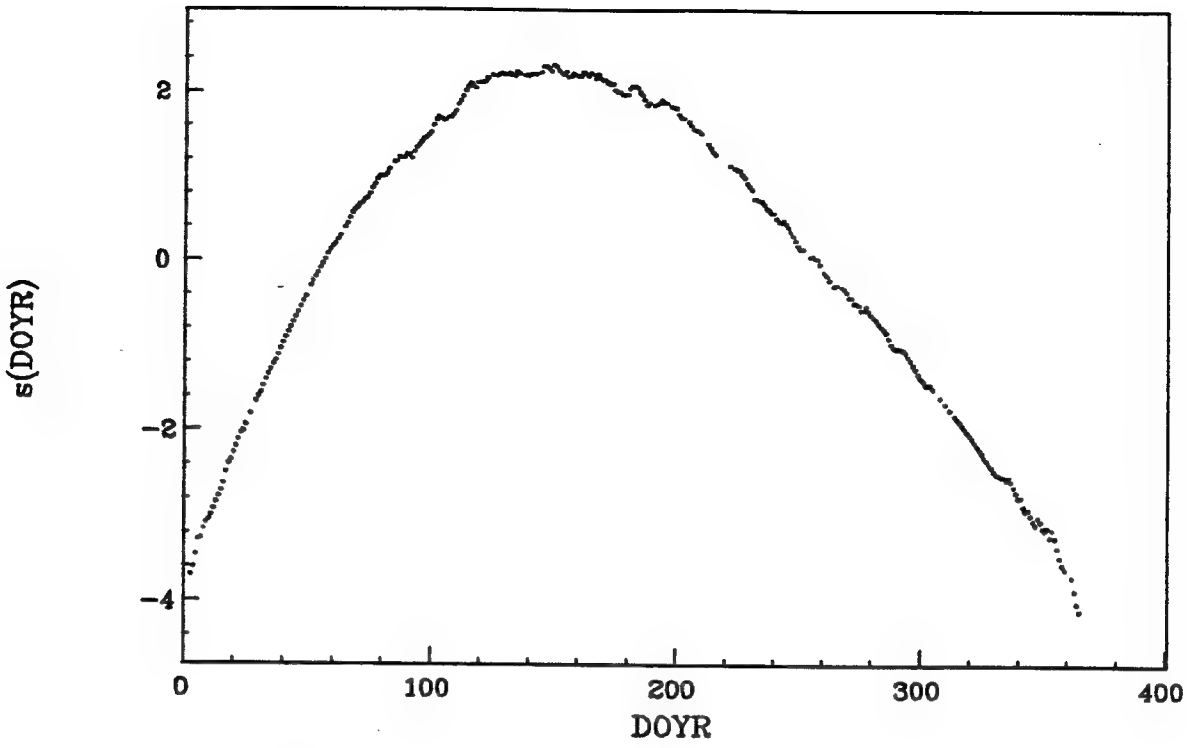


Figure 10. The estimated function for *Day of the Year*.

As a pre-screening step, we examined the effect of dropping each of the 8 covariates from the full model. For this phase, a fixed span of .5 was used. The full model had a residual sum of squares of 5176.56 on 22.5 degrees of freedom. Based on this, the estimate of residual error is 16.85. This compares to a standard multiple linear regression fit having a residual sum of squares of 6539.00 on 9 degrees of freedom. Using F-values as a rough guide, the variables SBTP, IBHT, DGPG and VSTY were seen to cause a significant increase in the residual sum of squares, and the remaining predictors were dropped.

Using these selected predictors, an additive model was fit, this time allowing the procedure to choose the optimal spans by cross-validation. The resultant model had a residual sum of squares of 5481.93 on 8.8 degrees of freedom, and is not significantly worse than the full model. We then added the variable DOYR to the model, and it was highly significant, decreasing the residual sum of squares by over 700, with 12.4-11.0=1.4 additional degrees of freedom. Note that instead of using DOYR's degrees of freedom (which was 1.9), we use the difference in degrees of freedom in the two models. The reason is that when a variable is added to a model, the spans chosen for other variables can change and hence their degrees of freedom change also.

Dropping any of these variables caused a large increase in the residual sum of squares. The fit of the full model (all 9 predictors) is also shown in Table 4.

At this point we mention a complexity caused by the varying spans. Although it seems

that the 4 variable model is 'nested" within the 8 variable model, there is no guarantee that its residual sum of squares will be higher. This is because the spans aren't chosen to minimize the overall residual sum of squares. It is quite possible that in dropping a useless predictor, the residual sum of squares will decrease! This is not a practical problem, but it would create some unusual twists in a testing theory for additive models.

Figures 6 – 10 show the estimated smooths for the 5 predictors. All the variables seem to display non-linear effects. To simplify the model, we forced in a parametric form for each of the variables, one at a time. A linear term was tried for SBTP and VSTY, and quadratic terms for IBHT, DGPG and DOYR. Both linear terms proved to be inadequate, especially for SBTP. Its linear term caused the residual sum of squares to increase to 5214.03, with about 1 less degree of freedom. The quadratic terms were all satisfactory. Thus we have our final model:

$$UPO3 = constant + s_1(SBTP) + a_1 IBHT + a_2 IBHT^2 + b_1 DGPG + b_2 (DGPG)^2 + s_3(VSTY) + c_1 DOYR + c_2 (DOYR)^2 \quad (37)$$

with a residual sum of squares of 4858.99 on 11.2 degrees of freedom. The RSS of this model is 112 higher than the RSS for the nonparametric model, but for descriptive purposes it is adequate.

Breiman and Friedman fit an ACE (Alternating Conditional Expectation) model to these data. The ACE model is the same as the additive Gaussian model, except that they also estimate a tranformation for the response. The final model obtained by Breiman and Friedman, using a forward stepwise ACE model contains the same predictors as the above model. In addition, the estimated transformation obtained from ACE was only slightly non-linear, and hence the estimated smooths from ACE are very similar to those in Figures 6 – 10.

We discuss an alternative to response variable transformation and apply to these data in Section 12.

12. Transformations of the Additive Model.

For exponential families, the model we have discussed up to now has the form $g(\mu) = \eta = \sum_1^p s_j(X_j)$, where $g(\cdot)$ is the (known) link function. A more general model is $g(\mu) = f(\eta) = f(\sum_1^p s_j(X_j))$, where $f(\cdot)$ is a unspecified smooth function. As we did for the $s_j(\cdot)$'s, we will show how to estimate $f(\cdot)$ non-parametrically. For ease of interpretation, we will restrict $f(\cdot)$ to be monotone. Note that since $f(\cdot)$ is arbritrary, so is $f^{-1}(g(\cdot))$, and we could write this as $g^*(\eta)$. Hence non-parametric estimation of $f(\cdot)$ provides non-parametric estimation of the link function. In some applications, we could set $g(\cdot)$ equal to the identity function; in others, we might want to start with $g(\cdot)$ equal to the natural link for the problem, and see if the estimated $f(\cdot)$ is close to the identity function.

Estimation of $f(\cdot)$ can be achieved through a modification of the local scoring algorithm. The new procedure consists of two alternating loops, one each for the estimation of the $s_j(\cdot)$'s

and $f(\cdot)$. To estimate $s(\cdot)$ with $\hat{f}(\cdot)$ fixed, we use formula (20) noting that the derivatives now involve the derivatives of $f(\cdot)$. In the generalized linear model case this reduces to

$$\eta^1(x) = \textbf{Smooth}[\eta^0(x) + (y - \mu^0)\frac{dg(\mu)}{d\mu^0})(\frac{1}{f'(\eta^0)})] \tag{38}$$

with weights $W^{-1} = [(\frac{dg(\mu)}{d\mu^0})^2/f'(\eta^0)^2]V^0$. To estimate $f(\cdot)$ with $\hat{s}(\cdot)$ fixed, let $b = s(x)$, and the update is

$$f^1(b) = \textbf{Smooth}\left[f^0(b) - \frac{\frac{dl}{d\eta^0}}{\textbf{Smooth}[\frac{d^2l}{d\eta^{0^2}}]}\right] \tag{39}$$

In the generalized linear models case, this is

$$f^1(b) = \textbf{Smooth}[f^0(b) + (y - \mu^0)\frac{dg(\mu)}{d\mu^0}] \tag{40}$$

with weights $W^{-1} = (\frac{dg(\mu)}{d\mu^0})^2 V^0$. The two loops, estimating the $s_j(\cdot)$'s and estimating $f(\cdot)$, are alternated until convergence.

For non-exponential family models, the link relation would be of the form $g(\theta) = f(\eta)$, where θ is some parameter of the likelihood. For example, one generalization of the Cox model would be $\lambda(t \mid x) = \lambda_0(t)e^{f(\sum_1^p s_j(x_j))}$.

The algorithm requires two special subroutines. First, the derivatives of $f(\cdot)$ are needed for the first step— Jerome Friedman kindly supplied us with his procedure for estimating the derivatives of a smooth. Secondly, the estimate of $f(\cdot)$ must be monotone. The monotone smoothing technique described in Friedman and Tibshirani (1984) was used for this purpose.

We tried this procedure on two examples. The first data set concerns the strength of yarns and is taken from Box and Cox (1964). It consists of a $3 \times 3 \times 3$ experiment, the response being number of cycles to failure and three covariates: length of test specimen, amplitude of loading cycle and load. Box and Cox fit linear terms to the covariates and found that the log transformation was ideal for the response. Note that if Y has mean μ and variance $V(\mu)$, then $\log Y$ has mean about $\log \mu$ and variance about $V(\mu)/\mu^2$. Hence if the log transformation is appropriate (both for additivity of effects and variance stabilization) then a generalized linear model with $\eta = \log(\mu)$ and $V(\mu) = \mu^2$ should also be appropriate.

To test our procedure, then, we set $g(\mu) = \mu$ and $V(\mu) = \mu^2$, and a linear term was fit for each covariate. The estimated $f(\cdot)$ is shown in figure 11, along with the exponential function. The agreement is very good. Furthermore, the residual sum of squares of the final model was very close to that obtained by Box and Cox for the log transformation.

As a second example, the link estimation procedure was applied to the ozone concentration data. Using the covariates of Table 4 and the spans chosen there, the estimated $f(\cdot)$ is shown in figure 12. The transformation has positive curvature, indicating that mild transformation of the additive predictor may improve the fit. This is in qualitative agreement with Breiman

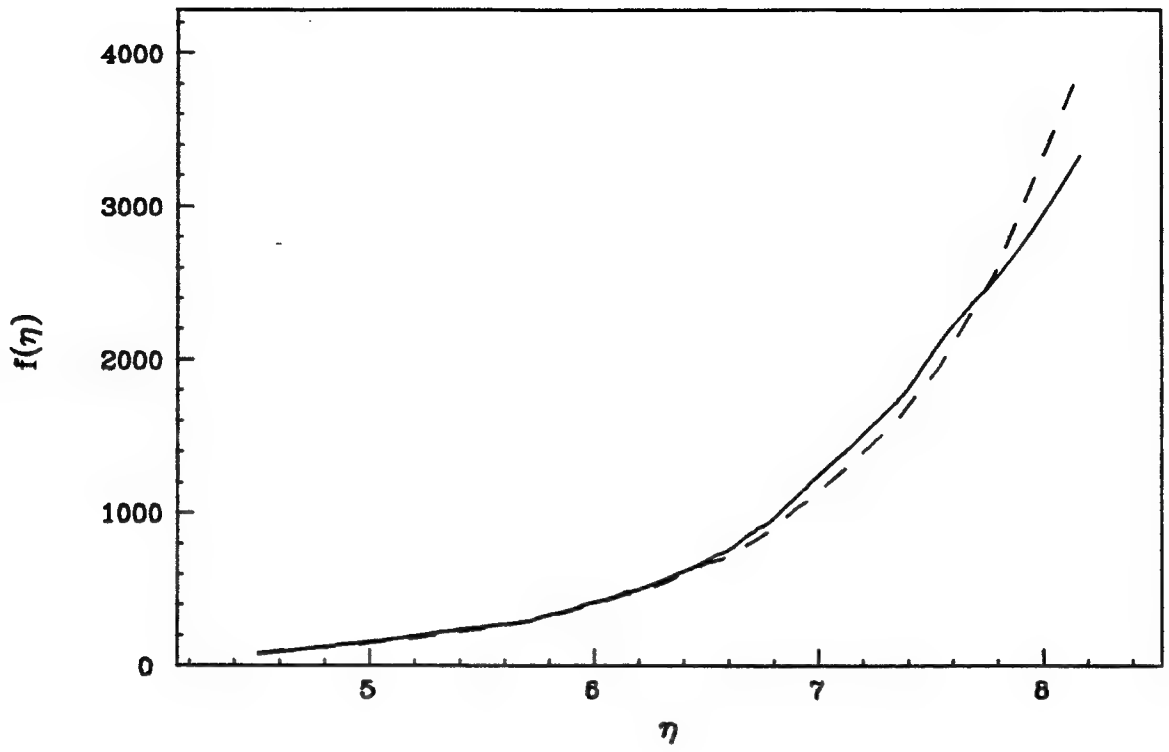


Figure 11. The solid function is the estimated transformation $f(\cdot)$ found for the Box-Cox data. The dashed function is the exponential suggested by the log-transform of the response variable y.

and Friedman's analysis: using a response variable transformation method (see ACE, next section), they find that a transformation with slightly negative curvature is appropriate.

Despite these encouraging results, the link estimation procedure is still experimental. The reason is that the iterations can be unstable if the derivatives of $f(\cdot)$ get too close to zero. We are presently studying this problem.

13. Relationship to Other Methods.

As we have seen, the Local Scoring technique generalizes standard (linear) likelihood methods. When each neighbourhood contains all the data points, Local Scoring corresponds exactly to standard (linear) maximum likelihood estimation. For smaller spans, the Local Scoring procedure can detect curvature in the covariate functions. In the class of GLM's, the linear predictor is generalized to a sum of smooth functions. Note also that this technique provides non-parametric "quasi-likelihood" estimation. McCullagh (1983) notes that in the exponential family, the score equation involves only the mean and variance of Y. This if one is only willing to assume a mean-variance relationship for Y, it might be reasonable to base the estimation procedure on the corresponding exponential family score equation. McCullagh calls this "quasi-

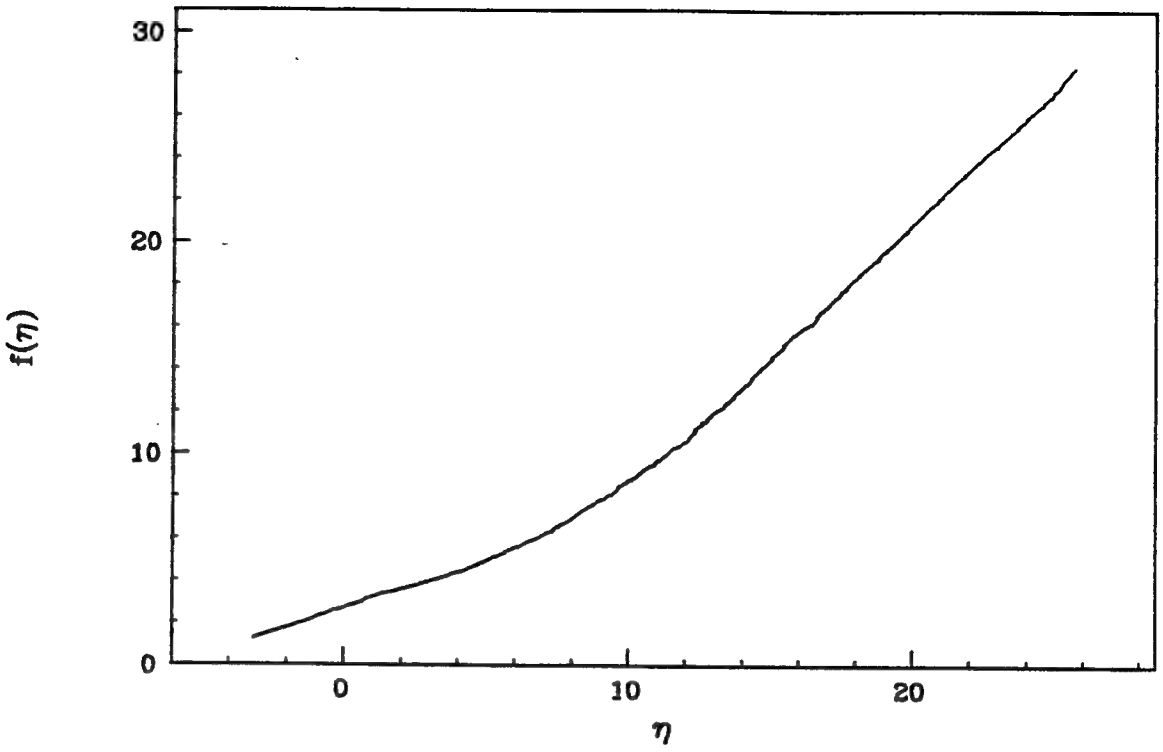


Figure 12. The estimated transformation for the Ozone Concentration Data of Section 11.

likelihood" estimation. Since local scoring (and local likelihood) depend only on the mean and variance as well, they can be thought of as extensions of quasi-likelihood modelling.

The models described in this paper are also related (when the error distribution is Gaussian) to a number of non-parametric regression models that have been recently suggested. Friedman and Stuetzle (1981) introduced the projection pursuit regression model:

$$Y = \sum_{1}^{p} s_j(a_j' X_j) + error \tag{41}$$

The directions a_j are found by a numerical search, while the $s_j(\cdot)'s$ are estimated by smoothers. Friedman and Stuetzle call the special case of co-ordinate directions, i.e. the model

$$Y = \sum_{1}^{p} s_j(X_j) + error \tag{42}$$

the "projection selection" model. This corresponds to the additive Gaussian model described here, and the algorithm for estimating the smooths is identical.

The ACE (Alternating Conditional Expectation) model generalizes the additive Gaussian

model by estimating a transformation of the response:

$$\theta(Y) = \sum_1^p s_j(X_j) + error \tag{43}$$

Related to this is the PACE (Predictive ACE) model of Friedman and Owen (1984)

$$Y = f(\sum_1^p s(X_i)) + error \tag{44}$$

This model transforms the mean of Y instead of Y. The PACE model is a special case of the the additive model with link estimation desribed in the previous section: it corresponds to the identity link ($\mu = \eta$) and Gaussian likelihood.

Finally, we note that the link estimation procedure of Section 12 can be used to further generalize the model to allow covariate projections (as in PPR). The model would be

$$g(\mu) = \sum_1^p \phi_j(a_j' x_j) \tag{45}$$

A single term $\phi(a'x)$ could be estimated by setting $f(\cdot) = \phi(\cdot)$ and forcing $s_i(x_i)$ to be linear ($= a_i x_i$). Additional terms could be added in a forward stepwise manner.

Software

A fortran program that computes local scoring estimates for exponential family models is available from either author.

Acknowledgements

We would like to thank Arthur Owen for his ideas on the expected log-likelihood criterion and for valuable discussions, Leo Breiman for allowing us to use his air pollution data, Werner Stuetzle for his proof in Appendix A, and Jerome Friedman for the use of his subroutine for calculating the derivatives of a smooth function.

References

Baker, R.J. and Nelder, J.A. (1978), *The GLIM System-Release 3.* Distributed by the Numerical Algorithms Group: Oxford.

Breiman, L. and Friedman, J.H. (1982), *Estimating Optimal Transformations for Multiple Regression and Correlation,* Dept. of Statistics Tech. Rept, Orion 16, Stanford University.

Cleveland, W.S. (1979), *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association, **74**, 829-836.

Cox, D.R. (1970), ***Analysis of Binary Data*, London: Chapman and Hall.**

Cox, D.R. (1972), *Regression models and life tables.* J. Roy. Stat. Soc. B, 34, 187-202.

Cox, D.R. (1975), *Partial likelihood.* Biometrika 62, 269-276.

Crowley, J. and Hu, M.(1977), *Covariance Analysis of heart transplant survival data.* J. Amer. Statist. Assoc. 72, 27-36.

Friedman, J.H. and Owen, A. Predictive ACE. (in preparation).

Friedman, J.H. and Tibshirani, R.(1984), *The Monotone Smoothing of Scatterplots.* Technometrics, August 1984.

Friedman, J.H. and Stuetzle, W. (1982), *Smoothing of Scatterplots,* Dept. of Statistics Tech. Rept. Orion 3, Stanford University.

Gong, G. (1984) Discussion in "Landwehr et. al. (1984)".

Haberman, S.J. (1976), *Generalized Residuals for Log-Linear Models.* Proc. 9^{th} Int'l Biometrics Conference, Boston, 104-122.

Hastie, T. (1983a), *Non-Parametric Logistic Regression.* Technical report ORION 16, Statistics Dept, Stanford University.

Hastie, T. (1983b), *Principal Curves.* Technical report ORION 24, Statistics Dept, Stanford University.

Hastie, T. (1984a) Discussion in "Landwehr et al (1984)".

Hastie, T. (1984b), ***Principal Curves and Surfaces,*** Stanford technical report and unpublished Phd thesis, Stanford University.

Kalbfleisch, J.D., and Prentice, R.L. (1980), *The statistical analysis of failure time data.* Wiley, New York.

Landwehr, J.M., Pregibon, D and Shoemaker, A.C. (1982), *Graphical Methods for assesing Logistic Regression Models.* J. Amer. Stat. Assoc., 79, 61-63.

McCullagh, P. (1983), *Quasi-Likelihood Functions.* Ann. Stat. 11, 59-67.

McCullagh, P. and Nelder, J. (1983), *Generalized Linear Models.* Chapman Hall, London.

Miller, R.G. and Halpern, J. (1981), *Regression with censored data.* Stanford Univ. technical report 66.

Nelder, J.A. and Wedderburn, R.W.M. (1972), *Generalized Linear Models.* J. Royal Statist. Soc. A **135**, 370-384.

Owen, A. (1983), *The Estimation of Smooth Curves.* (unpublished manuscript).

Stone, M. (1977), *An Asymptotic choice of Model by Cross-validation and Akaike's Criterion,* J. Roy. Stat. Soc. B, **7**, 44-47.

Tibshirani, R. (1982), *Non-Parametric Estimation of Relative Risk.* Technical Report ORION 22, Statistics Dept., Stanford University.

Tibshirani, R. (1984), *Local Likelihood Estimation.* Stanford technical report and unpublished Phd thesis, Stanford University.

Young, F.W, Takane, Y, and de Leuuw, J. (1978), *The Principal Components of Mixed Measurement Level Multivariate Data: an Alternating Least Squares Method with Optimal Scaling Features,* Psychometrika, **43**, no.2.

Appendix A. The backfitting algorithm using linear fits.

In this appendix we prove that the fit vector from the backfitting algorithm converges to the least squares answer if global linear fits are used to estimate $\mathbf{E}(\cdot \mid \boldsymbol{x})$. Let V be the subspace spanned by $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ with orthogonal complement $V^{\perp}$; here $\boldsymbol{x}_j$ is a n vector of observations on variable j. Let P_j denote linear projections onto $\boldsymbol{x}_j$. For any vector $\boldsymbol{a}$ let $\hat{\boldsymbol{a}}$ be its projection onto V and $\boldsymbol{a}^{\perp} = \boldsymbol{a} - \hat{\boldsymbol{a}}$. The residual after m cycles of the backfitting algorithm through the p predictors is given by

$$\begin{aligned} &\boldsymbol{r}^m = C^m \boldsymbol{y} \\ &\text{where} \\ &C = (I - P_p)(I - P_{p-1} \ldots (I - P_1)\boldsymbol{y}. \end{aligned} \tag{46}$$

It is immediate that $\boldsymbol{r}^m = C^m \hat{\boldsymbol{y}} + \boldsymbol{y}^{\perp}$.

Theorem 1

$\|C^m \hat{\boldsymbol{y}}\| \to 0$,and thus $\boldsymbol{r}^m \to \boldsymbol{y}^{\perp}$.

Proof. (Stuetzle, 1983)

For any vector $\boldsymbol{a}$ we use the natural norm for matrices to get

$$\begin{aligned} \|C\boldsymbol{a}\| &\le \|I - P_p\| \, \|(I - P_{p-1}) \ldots (I - P_1)\boldsymbol{a}\| \\ &\le \|(I - P_{p-1}) \ldots (I - P_1)\boldsymbol{a}\| \\ &\vdots \\ &\le \|\boldsymbol{a}\| \end{aligned} \tag{47}$$

since $\|I - P_j\| = 1 \; \forall \, j$. Similarly

$$\begin{aligned} \|C\boldsymbol{a}\| = \|\boldsymbol{a}\| \\ &\Rightarrow \|(I - P_1)\boldsymbol{a}\| = \|\boldsymbol{a}\| \\ &\Rightarrow (I - P_1)\boldsymbol{a} = \boldsymbol{a} \\ &\Rightarrow \boldsymbol{a} \in \boldsymbol{x}_1^{\perp} \end{aligned}$$

But then $\boldsymbol{a} \in \boldsymbol{x}_2^{\perp}$ .... and finally $\boldsymbol{a} \in \boldsymbol{x}_p^{\perp}$. Thus $\|C\boldsymbol{a}\| = \|\boldsymbol{a}\| \Rightarrow \boldsymbol{a} \in V^{\perp}$. So if $\boldsymbol{a} \in V$ then

$$\begin{aligned} \|C\boldsymbol{a}\| &< \|\boldsymbol{a}\| \\ &\le (1 - \epsilon) \, \|\boldsymbol{a}\| \, . \end{aligned}$$

where $0 < \epsilon \leq 1$. Also $C\boldsymbol{a} \in V$ for $\boldsymbol{a} \in V$. Hence

$$\begin{aligned} \|C^m \boldsymbol{a}\| &\leq (1-\epsilon) \|C^{m-1}\boldsymbol{a}\| \\ &\vdots \\ &\leq (1-\epsilon)^m \|\boldsymbol{a}\| \end{aligned}$$

This is true for any $\boldsymbol{a} \in V$. Since $\hat{\boldsymbol{y}} \in V$, the theorem is proved.

Appendix B. The use of weights in the backfitting procedure.

The function $f(x)$ that minimizes the weighted least squares problem $\mathbf{E}\, w(X)(Y - f(X))^2$ is simply $\mathbf{E}(Y \mid x)$, and the weights play no role. When we estimate this quantity using global or local straight line fits, however, the Gauss-Markov theorem tells us that we can estimate the parameters more efficiently by using weights.

If we wish to minimize

$$\mathbf{E}\, w(\boldsymbol{X})(Y - \sum s_j(X_j))^2 \tag{48}$$

and, as in the backfitting procedure, $s_1, \ldots, s_{p-1}$ are known, the situation is different. We write (48) as

$$\mathbf{E}_{X_p}\, \mathbf{E}_{X_1,\ldots,X_p,Y \mid X_p} \left[(R_p(Y, \boldsymbol{X}) - s_p(X_p))^2 \mid x_p \right] \tag{49}$$

where $R_p(Y, \boldsymbol{X}) = Y - \sum_{j \neq p} s_j(X_j)$. Minimizing this function w.r.t. s_p yields

$$\hat{s}_p = \frac{\mathbf{E}[w(\boldsymbol{X}) R_p(Y, \boldsymbol{X}) \mid x_p]}{\mathbf{E}[w(\boldsymbol{X}) \mid x_p]}. \tag{50}$$

Thus even in the distribution case, we need to use weights. It is clear that the weights play no role in the distribution case if they depend only on the variable on which we are conditioning.

In the generalized additive model situation, each iteration of the scoring procedure corresponds to a weighted least squares problem, with weights as specified in the algorithm. The weights, of course, usually depend on the unknown model and so the latest estimates are substituted.

Appendix C. Asymptotic equivalence of Local Scoring and Local Likelihood.

In this appendix, we sketch a proof that in the exponential family, the local likelihood estimate at a single X value asymptotically satisfies the local scoring update equation (19).

We assume that the Y_i's are independent with density $f_Y(y; \theta, \phi) = exp\{[y\theta - b(\theta)]/a(\phi) + c(y, \phi)\}$. We have $\mathbf{E}Y = b'(\theta) = \mu$, $VarY = b''(\theta) = V(\mu)\phi$, and μ is linked to a a single

covariate X by $\eta = s(x) = g(\mu)$. Consider estimation of $s(\cdot)$ at a fixed point x_0, with a sample size of k_n points in the neighbourhood N_0^n, and assume that $k_n \to \infty$ but

$$max_{\{i,j \in N_0^n\}} |x_i - x_j| = o(k_n^{-1/2}). \tag{51}$$

We make the simplifying assumption that the x_i's are equally spaced, hence the local likelihood estimate in the jth neighbourhood is just $\mu_j = \bar{y}_j$. The local scoring step is $\eta^1 =$ Smooth$[\eta + (y-\mu)\frac{d\eta}{d\mu}]$, with weights $(\frac{d\mu}{d\eta})^2 V(\mu)$. Assuming that $g(\mu)$ and $b(\theta)$ have two bounded derivatives, we can expand each in a Taylor Series, and using (51) , the local scoring step can be written as

$$\eta_0^1 - \ \mathbf{Ave}_0(\eta_j^0) + \ \mathbf{Ave}_0(y_j - \mu_j) + o(1) = 0 \tag{52}$$

where $\mathbf{Ave}_0$ represents the (unweighted) mean over $j \in N_0^n$. Thus it is sufficient to show that $\mu_j = \bar{y}_j$ and $\eta_j = g(\bar{y}_j)$ satisfy $\mathbf{Ave}_0(y_j - \mu_j) = o(1)$ and $\eta_0 - \ \mathbf{Ave}_0(\eta_j) = o(1)$, respectively. Assuming that the second moment of Y is well-behaved enough to allow application of the weak law, each of these follow by standard Taylor series arguments.